Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Nikolaj Janko

Algoritmično podprta optimizacija pospeševanja prodaje

magistrsko delo

magistrski program druge stopnje
računalništvo in informatika

Mentor: doc. dr. Jurij Mihelič

Ljubljana, 2017

Zahvala

Iskreno se zahvaljujem mentorju doc. dr. Juriju Miheliču za sodelovanje, nasvete, podporo in pomoč.
Zahvaljujem se tudi vsem ostalim, ki so na kakršen koli način prispevali k nastanku magistrskega dela.
Posebno se zahvaljujem domačim in bližnjim, ki so mi skozi ves študij stali ob strani.

Nikolaj Janko, 2017

Kazalo

Seznam uporabljenih kratic

kratica	angleško	slovensko
KNN	k-nearest neighbour	k najbližjih sosedov
RF	random forest	naključni gozdovi
RAM	random access memory	pomnilnik z naključnim dostopom
ASIN	amazon standard identification number	amazonova standardna identifikacijska številka
API	application programming interface	aplikacijski programski vmesnik

Povzetek

Naslov: Algoritmično podprta optimizacija pospeševanja prodaje

Spletna prodaja iz leta v leto strmo narašča, kar predstavlja dodaten izziv za prodajalce na spletnih prodajnih mestih, kot je Amazon, saj je konkurenčnih prodajalcev vedno več. Zaradi tega želi prodajalec ustvariti trdno vez med njim in kupcem, kar zahteva previdnost pri izbiri kupcev za promocijske aktivnosti, ki pa so za uspešno prodajo neizogibne. V magistrskem delu smo omenjeni problem definirali z modelom napovedovanja in zanj razvili več pristopov, ki temeljijo na teoriji grafov in množic. Zanje smo v sklopu ovrednotenja rezultatov testiranja na resničnih podatkih prodaje iz spletnega prodajnega mesta podjetja Amazon pokazali, da so v primerih napovedovanja kupcev boljša od regresijskih metod, ki so sicer ene od najpogosteje uporabljenih metod v napovedni analitiki.

Ključne besede

optimizacija promocij prodaje, amazon, algoritmi prodaje, napovedna analitika

Abstract

Title: Algorithmic optimization of sales acceleration

Online sales are growing rapidly over the years, which poses an additional challenge for online sellers such as Amazon, as there is ever more competition. The seller therefore wants to create a strong bond between him and the buyer. This requires caution in choosing buyers for promotional activities which are inevitable for successful sales. In the thesis we defined the problem with prediction model and developed several approaches based on graph theory and set theory to solve the problem. As part of the evaluation of test results on real sales data from Amazon online shopping site, we have shown that in case of customer prediction, the approaches we had developed are better than regression methods, the most commonly used methods in predictive analytics.

Keywords

sales promotion optimization, amazon, sales algorithm, predictive analytics

Poglavje 1

Uvod

Koncept spletne prodaje se je pojavil, še preden je Tim Berners-Lee izumil svetovni splet. Začelo se je leta 1979, ko je Michael Aldrich predstavil koncept "teleshopping"(danes znan kot spletna prodaja) na enem izmed prvih sistemov za izmenjavo tekstovnih sporočil, imenovanem Videotex. Leta 1994 je bilo ustanovljeno podjetje Amazon, ki je začelo s spletno prodajo knjig, danes pa je z več kot 3 024 000 sklenjenimi naročili na dan največje spletno prodajno mesto izdelkov iz več kot trideset različnih kategorij [1].

1.1 Motivacija in cilj

Prodajalci so skozi zgodovino spletne prodaje razvili tehnike, ki jih uporabljajo za povečanje števila naročil oziroma pospeševanje prodaje. Nekaj teh tehnik bo podrobneje predstavljenih v prvem delu magistrskega dela. Pri večini aktivnosti pospeševanja prodaje se prodajalec srečuje z vprašanjem: Kateri kupci? Primer so promocijske aktivnosti, kjer prodajalec želi nek izdelek, ki ga prodaja, čim bolj približati kupcem. Tedaj prodajalcu ni smiselno, da v promocijo vključi vse kupce, ki jih ima na voljo, saj bi s tem povečal stroške, poleg tega pa se večina kupcev, ki jih promovirani izdelek ne zanima, ne bi odzvala pozitivno, kar lahko privede tudi do izgube kupcev. Rešitev je, da prodajalec v promocijsko aktivnost vključi samo kupce, za

katere predvideva, da se bodo odzvali pozitivno, in s tem poveča uspešnost promocije ter zaupanje kupcev. Potencialne kupce lahko prodajalec izbere glede na interese kupcev, če te pozna. Vendar obstaja še drugi način, ki s pomočjo posebne analize iz zgodovine obnašanja napove prihodnje obnašanje kupcev. Tej analizi rečemo napovedna analitika in je pogosto uporabljena v oglaševanju in promocijah prodaje.

V sklopu magistrskega dela smo se osredotočili na dva problema, in sicer problem napovedovanja potencialnih kupcev za poljuben izdelek in problem napovedovanja izdelkov, ki potencialno ustrezajo poljubnemu kupcu. Glavni cilj je iz preteklih resničnih podatkov naročil tvoriti potencialno hitrejšo in bolj natančno napoved od do sedaj najpogosteje uporabljenih metod, kjer je glavna značilka relacija med kupcem in izdelkom naročila.

1.2 Pregled vsebine

Problem napovedovanja smo s pomočjo teorije množic najprej definirali v splošnem modelu, ki predstavi osnoven koncept napovedovanja, tega pa potem uporabili za definicijo bolj specifičnih modelov napovedovanja kupcev in napovedovanja izdelkov. Poleg teh dveh modelov smo definirali ocenjevalne kriterije za ovrednotenje uspešnosti napovedovanja, ki smo jih uporabili v eksperimentalnem ovrednotenju pristopov.

S pomočjo teorije grafov smo razvili tri pristope za napovedovanje kupcev in dva pristopa za napovedovanje izdelkov. Postopek smo za vsako uporabljeno podatkovno strukturo v pristopih natančno opisali in dopolnili z opisi algoritmov. Uvedli smo dva vhodna parametra, s katerima lahko do neke mere vplivamo na rezultate napovedovanja in ki naredita pristope bolj prilagodljive različnim virom in področjem uporabe.

Da smo razvito rešitev lahko preizkusili in ovrednotili, smo potrebovali nekaj podatkov o naročilih iz resničnega prodajnega okolja, kot je spletno prodajno mesto podjetja Amazon. Te smo prejeli v obliki podatkovne baze, ki nam je služila kot glavni vir. Podatke in strukturo podatkovne baze smo

kratko analizirali, obdelali in združili z dodatnimi podatki, ki smo jih pridobili s pomočjo API-vmesnika Amazona.

Uspešnost napovedovanja naših pristopov bomo predstavili v zadnjem delu magistrskega dela, kjer smo za primerjavo definirali nekaj bolj znanih metod, ki se jih običajno uporablja v napovedni analitiki. Te smo skupaj z našimi pristopi v sklopu testiranja nad podatki naročil implementirali ter rezultate analizirali, ovrednotili in primerjali med sabo.

"The aim of marketing is to know and understand the customer so well the product or service fits him and sells itself."

— Peter Drucker

Poglavje 2

Pregled področja

V tem razdelku bomo na kratko predstavili prodajo z uporabo spletnih in računalniških tehnologij oziroma spletno prodajo. Več o spletni prodaji, kot so osnovni pojmi in terminologija, bomo predstavili v prvem delu, medtem ko se bomo v drugem razdelku osredotočili na metode, ki jih prodajalci uporabljajo za pospeševanje prodaje. Na koncu bomo predstavili obstoječe postopke in metode, ki se uporabljajo za optimizacijo v različnih promocijah prodaje, in dva pojava, ki negativno vplivata na uspešnost prodajnih promocij.

2.1 Spletna prodaja

Poglejmo si nekaj osnovnih pojmov, ki ji srečujemo v svetu spletne prodaje in jih bomo uporabljali skozi magistrsko delo.

Prodajalec je oseba ali podjetje pod enim imenom (imenom trgovine ali znamke), ki v zameno za plačilo ponuja enega ali več izdelkov.

Kupec je oseba ali podjetje, ki je sklenilo vsaj eno naročilo izdelka z enim ali več prodajalci. Za kupca ni nujno, da je naročen izdelek že prejel.

Izdelek je lahko blago v fizični obliki, ki ga prodajalec ponuja in oglašuje. Izdelek je prav tako lahko storitev ali programska rešitev.

Naročilo je akcija, ki jo sproži kupec in povezuje vse tri najbolj pomembne člene prodaje (kupec, prodajalec, izdelek). Vsako naročilo zavezuje kupca in prodajalca do izmenjave plačila in izdelka oziroma izdelkov.

Prodajna promocija je postopek iskanja in prepričevanja potencialnega kupca v nakup izdelka, ki ga prodaja prodajalec. Poznamo več različnih prodajnih promocij, ki so naštete in opisane v razdelku 2.2. Promocija velja kot eden od aktivnosti pospeševanja prodaje.

Spletno prodajno mesto je spletna aplikacija, ki svojim uporabnikom nudi vsaj dve osnovni storitvi. Prvo storitev ponuja prodajalcem; to je storitev prodaje in oglaševanja izdelkov na njihovi spletni strani. Druga storitev je storitev nakupa, ki jo spletno prodajno mesto ponuja običajno širši skupini uporabnikom, kupcem. Nekatera spletna prodajna mesta ponujajo tudi storitve skladiščenja, pakiranja in pošiljanja izdelkov.

Različni prodajalci so skozi čas razvili več načinov, kako pospešiti oziroma povečati svojo spletno prodajo in s tem svoj dobiček. Eden od glavnih načinov so promocijske aktivnosti oziroma tržno komuniciranje. Promocijske aktivnosti ali na kratko promocije so aktivnosti, s katerimi prodajalec približa izdelek potencialnim kupcem. Oglaševanje, osebna prodaja, javni odnosi (publiciteta) in prodajne promocije so nekatere izmed najbolj znanih promocijskih aktivnosti [2]. Mi se bomo osredotočili predvsem na prodajne promocije, kjer je cilj prodajalca prepričati kupca v nakup izdelkov, ki jih prodajalec ponuja.

2.2 Prodajne promocije

V prodajne promocije uvrščamo več metod, ki jih v spletni prodaji uporabljajo prodajalci, da bi prepričali kupce v nakup njihovega izdelka. Nekaj najpogostejših tipov prodajnih promocij bomo na kratko povzeli v nadaljevanju [3, 4, 5].

Promocije s popusti

Popusti so dandanes ena najpogostejših promocijskih akcij. Pri popustu govorimo o akcijskem znižanju cene izdelka ali več izdelkov. Posamezen popust običajno velja za določen izdelek, lahko pa tudi za večjo skupino izdelkov, kadar govorimo o množičnih popustih (primer: -30 % na vse artikle). Popust lahko navadno izkoristi vsak kupec v obdobju aktivne promocije. Pomemben del promocij s popusti je oglaševanje, kjer prodajalec obvesti potencialne kupce o popustih.

Popuste prodajalci predstavljajo na več načinov. Običajno sta ob znižanem izdelku navedena znižana in polna cena izdelka. Poleg njiju je navadno navedena tudi količina popusta, največkrat v odstotkih znižanja, velikokrat pa tudi kot prihranek, ki je razlika med polno in znižano ceno izdelka.

Promocije s kuponi

Ideja tovrstne promocije je, da prodajalec kupcu v zameno za kupon nudi popust oziroma znižanje cene. Kupone kupec običajno prejme ob posebnih dogodkih ali programih, kot so: večkratna naročila, naročila nad določenim zneskom, posebne ponudbe, rojstni dnevi in programi zvestobe. Poznamo dva tipa kuponov. Najpogosteje uporabljeni kuponi so kuponi za enkratno rabo. Te lahko kupec unovči le enkrat. Drugi tipi kuponov pa so kuponi za množično rabo, kjer lahko posamezen kupon unovči več različnih kupcev, običajno v omejenem obdobju. Promocija s kuponi je dobra za ohranjanje obstoječih strank kot tudi za pridobivanje novih.

Kuponi v spletni prodaji so največkrat kratki nizi črk, ki jih prodajalec tvori na spletnem prodajnem mestu. Kupec te unovči tako, da jih ob naročilu vnese v spletni obrazec, kjer potem prejme popust.

Brezplačna dostava in vračilo

Pri dostavi izdelka pride do dodatnih stroškov, ki jih mora pogosto poravnati kupec. Sodeč po raziskavi Walker Sands [6], je 80 % bolj verjetno, da

kupec izbere izdelek, kjer mu tovrstnih stroškov ni treba plačati in jih namesto njega poravna sam prodajalec. Prav tako raje izbere izdelek tam, kjer ga v primeru poškodbe pri dostavi ali nezadovoljstvu lahko brezplačno vrne prodajalcu. Prodajalci običajno stroške dostave vključijo v ceno izdelka; tako kupec dobi občutek, da dostave ne plačuje on. Drug način je, da prodajalec nudi brezplačno dostavo za naročila z večjo količino izdelkov ali večjim končnim zneskom. Primer slednjega je ponudba brezplačne dostave za naročila v vrednosti več kot 50 evrov.

Kratka promocija

Pri kratki promociji (angl. flash sale) prodajalec nudi popust na določen izdelek za nek kratek omejen čas. S tem kupcu vzbudi občutek nuje po hitrem odzivu in nakupu izdelka. Kratka promocija je najbolj učinkovita v prvi uri, saj se takrat zgodi več kot 50 % prodaje [7].

Ugodnosti na količino

Kupcu lahko prodajalec nudi različne ugodnosti na količino. S tem poskrbi, da proda več zaloge in hkrati ugodi kupcu, saj le-ta dobi nekaj v zameno. Prodajalec kupcem navadno nudi popust na količino, lahko pa tudi gratis izdelek (izdelek, ki ga kupec prejme brezplačno). Tovrstnih promocij se običajno ne uporablja za dražje izdelke ali izdelke, ki niso porabljivi (televizijski sprejemnik, avtomobil, mobilni telefon ...).

Podaritve in darila

Podaritve ali darila so odlična promocija za nove izdelke. S podarjenimi izdelki prodajalec pokaže kakovost in uporabnost izdelka ter s tem širi prepoznavnost. Izdelke prodajalec največkrat podarja strankam, ki v zameno za podarjen izdelek širijo novico o izdelku preko socialnih omrežij, video posnetkov, forumov, člankov in pregledov (angl. reviews) na spletnih prodajnih mestih. Primer teh strank so lastniki YouTube kanalov, kjer odpirajo

in predstavljajo različne izdelke (angl. unboxing).

Program zvestobe

Pri gradnji močne blagovne znamke je pomembno, da pridobljene zveste stranke obdržimo. Zato je dobro, da vodimo evidenco stalnih strank in gradimo bazo zvestih kupcev. Zveste kupce lahko nagradimo s popusti, drugimi ekskluzivnimi promocijami ali pa z zbiranjem točk/pik, ki jih lahko unovčijo.

2.3 Optimizacija prodajnih promocij

Uspešnost prodajne promocije je odvisna od več dejavnikov, kot so: tip prodajne promocije, izbira ciljne skupine (kupcev), vizualna oblika prodajne promocije, čas in kraj izvajanja promocije itd. Za merjenje uspešnosti promocij v spletni prodaji prodajalci največkrat uporabljajo podatke o prodajah promoviranega izdelka [5].

Skozi izkušnje so prodajalci identificirali več pojavov, ki zavirajo uspešnost prodajnih promocij. Nekaj teh bomo našteli v nadaljevanju.

Spreminjanje interesov kupcev

Čas spreminja kupce in z njimi tudi njihove interese (angl. interest drifting), česar se mora prodajalec pri tvorjenju promocijskih akcij zavedati. Za orodja in metode, ki jih prodajalci uporabljajo pri tvorjenju prodajnih promocij, je torej priporočljivo, da upoštevajo pojav spreminjanja interesov s tem, da imajo novejši podatki večjo težo od preteklih. Pri metodah podatkovnega rudarjenja lahko v ta namen uvedemo funkcijo pozabljanja, ki vhodne podatke uteži glede na starost podatka. Če je podatek star, ga funkcija pozabljanja uteži z manjšo utežjo kot novejše podatke.

Primer, ki odpravlja pojav spreminjanja interesov kupcev, je funkcija pozabljanja, uporabljena za utežitev lastnosti kupcev [8].

Pojav AdFatigue

AdFatigue je pojav, ki je najbolj znan pri oglaševanju, vendar se pojavlja tudi pri različnih prodajnih promocijah.

AdFatigue je pojav, ko ciljni kupec izgubi zanimanje za oglase ali promocije, ki so mu ponujene. Do tega lahko pride zaradi več razlogov, kot so: promocije izdelkov, ki zanj niso zanimive, prepogoste promocije, lažne ali zavajajoče promocije. Posledica je, da ciljni kupec začne promocije ignorirati ali pa se odjavi z elektronskega poštnega seznama.

Pojav kanibalizacije

Pojav kanibalizacije v promocijskem svetu je, ko interesi za prvi izdelek, ki ga ponuja prodajalec, izpodrinejo interese drugega izdelka istega prodajalca. Ta pojav je najbolj kritičen pri izdelkih enakega namena ali variacijah istega izdelka.

Do pojava prihaja tudi pri prodajnih promocijah in se ga običajno rešuje s časovnimi zamiki le-teh. Promocije, ki jih prodajalec izvaja hkrati ali v kratkem časovnem obdobju, namreč lahko vplivajo na uspešnost druga druge [9].

Da bi maksimirali uspešnost in minimizirali vplive negativnih pojavov prodajne promocije, prodajalci uporabljajo različne mehanizme, kot so načrtovanje, strateške analize, merjenje uspešnost promocij in optimizacijske metode. Mi se bomo osredotočili predvsem na optimizacijske metode.

V nadaljevanju bomo predstavili nekaj orodij oziroma tehnik, s katerimi si prodajalci pomagajo pri tvorjenju optimalnih prodajnih promocij.

Segmentacija kupcev

Segmentacija kupcev je tehnika, s katero prodajalec kupce razdeli v segmente glede na njihove lastnosti. Lastnosti so lahko različne, kot na primer: interesi, hobiji, zahteve, karakteristike, poklic, starost, religija, geografske lastnosti in podobno.

Pri tvorjenju prodajnih promocij se segmentacija kupcev uporablja za tesnejše prilagajanje promocije kupcem. Po segmentaciji prodajalec za vsak segment pozna skupno lastnost kupcev, ki so v segmentu. Potemtakem se lahko prodajalec osredotoči na posamezen segment kupcev in zanje pripravi bolj zanimivo prodajno promocijo.

Segmentacijo pri prodajnih promocijah prodajalci običajno izvedejo v treh osnovnih korakih. V prvem koraku prodajalec izbere pomembnejše lastnosti ciljnih kupcev, ki najbolj ustrezajo prodajnim promocijam, ki jih želi optimizirati. V drugem koraku iz izbranih lastnostih prodajalec izlušči nekaj najbolj pomembnih. Vse kombinacije teh izbranih lastnosti so dobljeni segmenti. V tretjem koraku prodajalec za modele segmentov zbere podatke ter oceni ustreznost posameznega segmenta glede na velikost segmenta in lastnosti kupcev v njem. Zanima ga predvsem, ali je segment ustrezen za izvedbo prodajne promocije nad njim.

Obstaja še drugi način tvorjenja segmentov, in sicer s pomočjo podatkovnega rudarjenja, kjer kupec prepusti računalniškim metodam, da same poiščejo naravno pojavljajoče se segmente. Pri tem se najpogosteje uporabljajo metode gručenja (angl. clustering algorithms) in regresijske metode, kot je logistična regresija [10, 11].

Izogibanje pojavu AdFatigue

AdFatigue pojav lahko prodajalec minimizira s tem, da se pri tvorjenju promocij drži določenih načel, ki jih predlagajo marketinški svetovalci [12]:

Manjša pogostost promocij

Večina kupcev se strinja, da pogoste ponudbe hitro postanejo nadležne, zato je pomembno, da prodajalci pogostost promocij zmanjšajo.

Večja zanimivost/mamljivost ponudb

Ponudbe naj bodo skrbno zasnovane z večjimi ugodnosti, kar bolj pritegne kupca.

Vsebina bližja kupcem

To lahko prodajalec doseže s pomočjo že omenjene segmentacije kupcev, kjer kupcem ponujamo zanj potencialno bolj zanimive ponudbe. Drug način je, da prodajalec s pomočjo napovedne analitike sestavi profil vsakega kupca in ponudbe tvori za vsakega posameznika posebej. V pomoč je lahko tudi povzeti profil kupca, sestavljen iz skupine zvestih kupcev.

Spreminjanje oblike promocij

Spremembe privlačijo nove kupce in vzpodbudijo večje zanimanje zvestih kupcev. Zato je priporočljivo, da prodajalci menjajo tipe prodajnih promocij (kuponi, masovni popusti, brezplačne dostave, kupi enega – dobiš dva ...).

Sprememba predstavitve ponudb

Običajno se v spletni prodaji ponudbe pošiljajo preko elektronske pošte. Dobro je, da prodajalec poskuša čim večkrat menjati vizualno obliko ponudbe, kar bo vedno znova pritegnilo kupca.

Menjava ciljne skupine

Prodajalcu se ni potrebno popolnoma osredotočiti na določen segment kupcev. Včasih je dobro kako ponudbo poslati tudi kateri drugi skupini kupcev; mogoče bo odziv boljši, kot je bil predviden.

Filtriranje posameznih kupcev

Iz promocije je dobro izključiti kupce, ki so že kupili promovirani izdelek ali so že odprli podobno promocijo.

Napovedna analitika

Pri promocijah se vedno pojavljajo vprašanja, kot so: Ali se bo kupec pozitivno odzval? Kateri kupci imajo največji potencial? Kdaj je dobro izvesti promocijo? Na vsa ta vprašanja je lažje odgovoriti, če poznamo prihodnost. Prodajalci se dejansko ukvarjajo z napovedovanjem prihodnosti obnašanja

kupcev (angl. predictive analytics) [13]. Prav z omenjenim problemom se bomo v tem magistrskem delu ukvarjali tudi mi.

Pod napovedno analitiko uvrščamo različne metode, ki poskušajo iz analize preteklih podatkov ali preteklega vedenja napovedati prihodnje obnašanje. Za izvajanje tovrstne analitike v spletni prodaji se uporabljajo metode iz različnih matematičnih in računalniških smeri, med katerimi prevladujejo metode podatkovnega rudarjenja in teorija iger [14, 13].

Napovedna analitika je vedno bolj razširjen pojem v svetu spletne prodaje in v večjih marketinških podjetjih, ki se uporablja na več področjih marketinga. Prvi primer je uporaba napovedne analitike v oglaševanju, kjer jo prodajalci in analitiki uporabljajo za analizo in napovedovanje odzivov kupcev na različne oglase ob različnih časih. Kot drugi primer je uporaba napovedne analitike pri strateških odločanjih marketinga, kjer jo uporabljajo za izbiro strategij, izbiro cen izdelkov, zagon novih izdelkov, optimizacijo odločanja in optimizacijo marketinških ciljev. Poleg ostalih primerov uporabe bomo izpostavili še tretjega, to je uporaba v prodajnih promocijah. Pri teh se napovedna analitika uporablja za napovedovanje uspešnosti promocij, potencialnih kupcev, odzivov kupcev na promocije in mnogo več [15, 16].

"The best way to predict the future is to create it."

— Peter Drucker

Poglavje 3

Napovedovalni model

V tem poglavju bomo definirali dva problema napovedovanja za namen optimizacije promocij ali uporabo v drugih marketinških aktivnostih. Prvi bo problem napovedovanja kupcev za poljuben izdelek. Rešitev omenjenega napovednega problema bi pomagala prodajalcem pri promociji svojih izdelkov in zmanjšala vpliv pojava AdFatigue, ki se pojavi, ko kupec začne ignorirati promocije zaradi nezanimanja ali drugih razlogov.

Drugi problem je napovedovanje izdelkov za kupca. Rešitev tega problema bi omogočila tvorjenje promocij glede na interese kupca, kar bi pripomoglo predvsem promocijskim spletnim mestom (npr. Kuponko [17], Uberzonclub [18]), kjer se kupci prijavljajo z namenom, da prejmejo popuste ali kupone.

Najprej bomo definirali problem napovedovanja v splošnem ter pripravili napovedni model, ki ga bomo uporabili pri napovedovanju kupcev in napovedovanju izdelkov za razvoj bolj specifičnih napovednih modelov. V nadaljevanju se bomo osredotočili na osnovne definicije, kjer bomo definirali kupca in izdelek. Pojme bomo definirali s pomočjo teorije množic in relacij med njimi. Sledila bo definicija problema napovedovanja kupcev, za njo pa definicija problema napovedovanja izdelkov.

3.1 Napovedovanje v splošnem

Napovedovanje lahko predstavimo z napovednim modelom, ki je funkcija $\boldsymbol{f}(\boldsymbol{I_f})$, kjer so $\boldsymbol{I_f}$ vhodni parametri, na podlagi katerih funkcija $\boldsymbol{f}$ tvori izhod oziroma napoved. Za napovedovanje velja, da želimo napoved čim bolj približati resničnemu stanju, ki pa je izhod resničnostne funkcije $\boldsymbol{r}(\boldsymbol{I_r})$. Napoved je popolnoma pravilna, kadar napovedna funkcija vrne napoved, enako resničnemu stanju, oziroma: $\boldsymbol{f}(\boldsymbol{I_f}) = \boldsymbol{r}(\boldsymbol{I_r})$.

V praksi funkcije $\boldsymbol{r}$ ne poznamo v celoti, saj gre običajno za zelo kompleksne in zahtevne procese, ki jih funkcija $\boldsymbol{r}$ izvaja. Funkcija $\boldsymbol{r}$ je za nas torej neznanka. V vhodne parametre $\boldsymbol{I_f}$ pa želimo vključiti čim več vhodnih parametrov iz $\boldsymbol{I_r}$, ki nastopajo pri resničnostni funkciji $\boldsymbol{r}$. S tem model napovedovanja približamo resničnostni funkciji in poskrbimo, da pri napovedi uporabimo čim manj parametrov, ki v resnici ne vplivajo na izhod resničnostne funkcije $\boldsymbol{r}(\boldsymbol{I_r})$.

Ko tvorimo pristope za napovedne modele, jih običajno želimo oceniti. To storimo tako, da implementacijo funkcije $\boldsymbol{f}$ napovednega modela, ki ga ocenjujemo, uporabimo nad primerom, za katerega poznamo resnično stanje $\boldsymbol{r}(\boldsymbol{I_r})$. Vhod $\boldsymbol{I_r}$ resničnostne funkcije $\boldsymbol{r}$ pri napovedi uporabimo kot vhod $\boldsymbol{I_f}$ v napovedni model $\boldsymbol{f}$. Napoved $\boldsymbol{f}(\boldsymbol{I_f})$ napovednega modela $\boldsymbol{f}$ nato s funkcijo ocenjevanja, ki jo definiramo v postopku ocenjevanja, ocenimo tako, da napoved primerjamo z resničnim stanjem $\boldsymbol{r}(\boldsymbol{I_r})$. Resničnemu stanju rečemo tudi testna množica. Pri testiranju torej velja $\boldsymbol{I_r} = \boldsymbol{I_f}$.

Poglejmo si dva kratka primera modeliranja problemov:

Primer 1 (Napovedovanje vremena)**.** *V tem primeru želimo napovedati, kakšno bo vreme naslednji dan.*

Vhod $\boldsymbol{I_r}$ v resničnostno funkcijo $\boldsymbol{r}$ so vsi dejavniki, ki vplivajo na vreme in podnebje. Poglejmo si nekaj teh dejavnikov: letni čas, pretekla vremenska slika, moč sonca, metulj v Afriki ... Seveda je teh dejavnikov veliko, nekaterih sploh ne poznamo, za druge pa nimamo podatka.

Resničnostna funkcija $\boldsymbol{r}$ je kompleksen proces, ki ga izvaja narava. Izhod resničnostne funkcije $\boldsymbol{r}$ oziroma resnično stanje je vreme, ki ga prinese naslednji dan.

Sedaj lahko sestavimo primer modela napovedi. Za vhod $\boldsymbol{I_f}$ vzemimo tri dejavnike: vreme danes, letni čas in hitrost vetra. Za funkcijo $\boldsymbol{f}$ vzemimo napovedovanje z linearno regresijo. Izhod naše napovedi bo ena od vrednosti iz sledečega nabora: sončno, deževno, oblačno.

Napovedni model ocenimo na historičnih podatkih vremena. Za oceno lahko uporabimo preprosto primerjalno funkcijo enakost.

Primer 2 (Prepoznavanje predmeta na sliki)**.** *Cilj je, da napovemo predmet, prikazan na sliki.*

Vhod $\boldsymbol{I_r}$ v resničnostno funkcijo $\boldsymbol{r}$ je slika oziroma so svetlobni žarki, odbiti od površine slike. Resničnostna funkcija $\boldsymbol{r}$ je postopek prepoznave, ki se izvaja vse od očesa skozi možgane opazovalca. Izhod oziroma resnično stanje $\boldsymbol{r}(\boldsymbol{I_r})$ je pisna ali verbalna oblika besede, ki predstavlja prepoznan predmet na sliki.

Definirajmo še potencialen napovedni model. Za razliko od primera 1 tukaj natančno poznamo vhod v napovedni model $\boldsymbol{I_f}$, ki je digitalna oblika slike (polje številčnih vrednosti, ki predstavljajo vsako točko na sliki). Za funkcijo napovednega modela $\boldsymbol{f}$ vzamemo naučeno nevronsko mrežo za prepoznavo predmetov na sliki. Napoved napovedne funkcije je predmet, zapisan v elektronski tekstovni obliki (niz ASCII-znakov).

Napovedni model lahko ocenimo na primerih slik, na katerih je opazovalec že prepoznal predmet. Tudi tukaj lahko uporabimo enakost nizov za primerjalno funkcijo. Vendar bi bilo smiselno uporabiti kakšno naprednejšo ocenjevalno funkcijo, ki upošteva tudi hierarhijo predmetov.

3.2 Osnovne definicije

Iz podatkov o prodaji, ki jih imamo na voljo, lahko tvorimo dve osnovni množici. Prva množica je množica vseh kupcev, druga pa je množica vseh

izdelkov. Da se bomo lažje sklicevali na množice in njihove elemente, smo jim dodelili naslednje oznake:

- $\boldsymbol{\mathcal{B}}$ ... množica vseh kupcev,
- $\boldsymbol{\mathcal{I}}$ ... množica vseh izdelkov,
- $\boldsymbol{b} \in \boldsymbol{\mathcal{B}}$ ... posamezen kupec,
- $\boldsymbol{i} \in \boldsymbol{\mathcal{I}}$ ... posamezen izdelek.

Vsak kupec, ki bo nastopal v našem problemu, je vedno del množice vseh kupcev. Prav tako velja za vse izdelke, da so vedno del množice vseh izdelkov.

Pri napovedih se bomo osredotočili na interese kupcev. Za kupca $\boldsymbol{b}$ želimo napovedati množico izdelkov, ki jih bo kupec potencialno naročil. Zanj obstaja tudi množica izdelkov, ki jih je naročil, vendar pa je mi pri napovedovanju ne poznamo. Definirati moramo še dve množici, ki ju bomo uporabljali pri napovedi kupcev za izdelek. Množica $\boldsymbol{B_i}$ vsebuje vse kupce, ki so naročili izdelek $\boldsymbol{i}$. V napovedni množici $\boldsymbol{A_i}$ pa so kupci, za katere napovedujemo, da bodo kupili izdelek $\boldsymbol{i}$.

Omenjene množice bomo označili z naslednjimi oznakami:

- $\boldsymbol{B_i} \subseteq \boldsymbol{\mathcal{B}}$ ... množica kupcev, ki so naročili izdelek $\boldsymbol{i} \in \boldsymbol{\mathcal{I}}$,
- $\boldsymbol{A_i} \subseteq \boldsymbol{\mathcal{B}}$ ... množica napovedanih kupcev izdelka $\boldsymbol{i} \in \boldsymbol{\mathcal{I}}$,
- $\boldsymbol{I_b} \subseteq \boldsymbol{\mathcal{I}}$ ... množica izdelkov, ki jih je naročil kupec $\boldsymbol{b} \in \boldsymbol{\mathcal{B}}$,
- $\boldsymbol{C_b} \subseteq \boldsymbol{\mathcal{I}}$ ... množica za kupca $\boldsymbol{b} \in \boldsymbol{\mathcal{B}}$ napovedanih izdelkov.

Naslednji dejstvi opisujeta lastnost simetričnosti med kupci in izdelki, ki pa ju lahko združimo v posledico 3.2.1.

Dejstvo 3.2.1. Za vsak izdelek $\boldsymbol{i} \in \boldsymbol{I}$ velja: če ga je naročil kupec $\boldsymbol{b} \in \boldsymbol{\mathcal{B}}$, potem je $\boldsymbol{b}$ del množice kupcev $\boldsymbol{\mathcal{B}_i}$, ki so naročili izdelek $\boldsymbol{i}$, oziroma

$$\exists \boldsymbol{b} \in \boldsymbol{\mathcal{B}} : \boldsymbol{i} \in \boldsymbol{I_b} \implies \boldsymbol{b} \in \boldsymbol{B_i}.$$

Dejstvo 3.2.2. Za vsakega kupca $\boldsymbol{b} \in \boldsymbol{B}$ velja: če je naročil izdelek $\boldsymbol{i} \in \boldsymbol{\mathcal{I}}$, potem je $\boldsymbol{i}$ del množice izdelkov $\boldsymbol{\mathcal{I}_b}$, ki jih je kupec $\boldsymbol{b}$ kupil, oziroma

$$\exists \boldsymbol{i} \in \boldsymbol{\mathcal{I}} : \boldsymbol{b} \in \boldsymbol{B_i} \implies \boldsymbol{i} \in \boldsymbol{I_b}.$$

Iz obeh dejstev 3.2.1 in 3.2.2 sledi naslednja posledica:

Posledica 3.2.1.

$$\forall \boldsymbol{i} \in \boldsymbol{\mathcal{I}}, \boldsymbol{b} \in \boldsymbol{\mathcal{B}} : \boldsymbol{i} \in \boldsymbol{I_b} \iff \boldsymbol{b} \in \boldsymbol{B_i}$$

3.3 Napovedovanje kupcev

Cilj problema napovedovanja kupcev je, da za izdelek $\boldsymbol{i} \in \boldsymbol{\mathcal{I}}$ napovemo kupce, ki bodo z visoko verjetnostjo naročili izdelek $\boldsymbol{i}$. Za definiranje problema bomo uporabili napovedni model, ki smo ga definirali v prejšnjem razdelku 3.1.

Definicija resničnostne funkcije

Kot vhod $\boldsymbol{I_r}$ v resničnostno funkcijo $\boldsymbol{r}$ so vsi kupci, ki jim je na voljo izdelek $\boldsymbol{i}$, ter vsi dejavniki, ki vplivajo na odločitev teh kupcev, ali bodo izdelek $\boldsymbol{i}$ naročili ali ne. Omenjenih dejavnikov je zelo veliko, zato jih bomo tukaj našteli le nekaj:

- razpoložljivost izdelka $\boldsymbol{i}$,
- kako zelo si kupec želi izdelek $\boldsymbol{i}$,
- koliko je izdelek $\boldsymbol{i}$ znižan,
- cena izdelka $\boldsymbol{i}$,
- rang izdelka na spletnem mestu,
- stroški naročila izdelka,
- premožnost kupca,

- ali je izdelek $\boldsymbol{i}$ v promociji,
- trenutno emocionalno stanje kupca,
- ...

Resničnostna funkcija $\boldsymbol{r}$ je razmišljanje in obnašanje posamezna kupca v zvezi z izdelkom $\boldsymbol{i}$, kar pa velja za zelo kompleksno in nedeterministično.

Izhod modela so pozitivni in negativni odzivi kupcev na izdelek $\boldsymbol{i}$. Kot pozitiven odziv kupca velja nakup izdelka $\boldsymbol{i}$ ali pa samo zanimanje za izdelek $\boldsymbol{i}$. Za negativen odziv smatramo ignoriranje ali celo grajanje izdelka $\boldsymbol{i}$. Kupcem iz množice $\boldsymbol{B_i} \subseteq \boldsymbol{\mathcal{B}}$, ki so se za izdelek $\boldsymbol{i}$ odzvali pozitivno, bomo rekli tudi dejanski kupci.

Definicija napovednega modela

V nadaljevanju bomo natančno definirali enega od več možnih modelov napovedovanja kupcev za izdelek $\boldsymbol{i}$. Naš napovedni model bo prilagojen za napovedovanje dejanskih kupcev iz preteklih podatkov naročil.

Nabor vhodnih parametrov $\boldsymbol{I_f}$ našega napovedovalnega modela za izdelek $\boldsymbol{i}$ so podatki, ki povedo, katere izdelke so posamezni kupci $\boldsymbol{b} \in \boldsymbol{\mathcal{B}}$ kupovali v preteklosti. Napoved oziroma izhod napovednega modela je množica kupcev, za katere napovedujemo pozitiven odziv na izdelek $\boldsymbol{i}$. Slednjo izhodno množico kupcev bomo označevali z $\boldsymbol{A_i}$.

Napovedno funkcijo $\boldsymbol{f}$ lahko opišemo kot postopek, ki za izdelek $\boldsymbol{i} \in \boldsymbol{\mathcal{I}}$ tvori napovedno množico kupcev $\boldsymbol{A_i}$. Pri tvorjenju napovedne množice lahko funkcija $\boldsymbol{f}$ iz množice $\boldsymbol{\mathcal{B}}$ izbere kupce, ki so tudi del množice dejanskih kupcev $\boldsymbol{B_i}$. Če velja $\boldsymbol{b} \in \boldsymbol{A_i} \cap \boldsymbol{B_i}$, je za izdelek $\boldsymbol{i}$ napovedani kupec $\boldsymbol{b}$ tudi dejanski kupec, kar pomeni, da je kupec $\boldsymbol{b}$ napovedan pravilno.

Število kupcev v množici $\boldsymbol{A_i} \cap \boldsymbol{B_i}$ nam pove, koliko dejanskih kupcev izdelka $\boldsymbol{i}$ smo napovedali pravilno. Zato bomo omenjeno množico označili z $\boldsymbol{T^+}$ (angl. true positives).

Definicija 3.3.1. Množica pravilno napovedanih kupcev izdelka $\boldsymbol{i} \in \boldsymbol{\mathcal{I}}$:

$$\boldsymbol{T_i^+} = \boldsymbol{A_i} \cap \boldsymbol{B_i}.$$

Množico nepravilno napovedanih (angl. false positives) za izdelek $\boldsymbol{i}$ tvorijo tisti kupci, ki so del napovedne množice $\boldsymbol{A_i}$, vendar niso del množice pravilno napovedanih kupcev $\boldsymbol{T_i^+}$. Slednjo množico bomo označili z oznako $\boldsymbol{F_i^+}$.

Definicija 3.3.2. Množica nepravilno napovedanih kupcev izdelka $\boldsymbol{i} \in \boldsymbol{\mathcal{I}}$:

$$\boldsymbol{F_i^+} = \boldsymbol{A_i} \setminus \boldsymbol{T_i^+} = \boldsymbol{A_i} \setminus (\boldsymbol{A_i} \cap \boldsymbol{B_i}) = \boldsymbol{A_i} \setminus \boldsymbol{B_i}.$$

Množico dejanskih kupcev izdelka $\boldsymbol{i}$, ki smo jih pri napovedi spregledali, bomo označili z oznako $\boldsymbol{F_i^-}$ (angl. false negatives).

Definicija 3.3.3. Množica spregledanih dejanskih kupcev za izdelek $\boldsymbol{i} \in \boldsymbol{\mathcal{I}}$:

$$\boldsymbol{F_i^-} = \boldsymbol{B_i} \setminus \boldsymbol{T_i^+} = \boldsymbol{B_i} \setminus (\boldsymbol{A_i} \cap \boldsymbol{B_i}) = \boldsymbol{B_i} \setminus \boldsymbol{A_i}.$$

V četrto množico $\boldsymbol{T_i^-}$ (angl. true negatives) pa uvrščamo kupce, ki jih za izdelek $\boldsymbol{i}$ nismo napovedali in tudi niso dejanski kupci izdelka $\boldsymbol{i}$.

Definicija 3.3.4. Množica nenapovedanih kupcev, ki niso dejanski kupci izdelka $\boldsymbol{i} \in \boldsymbol{\mathcal{I}}$, je:

$$\boldsymbol{T_i^-} = (\boldsymbol{A_i} \cup \boldsymbol{B_i})^{\complement},$$

kjer $\boldsymbol{X}^{\complement}$ označuje komplement množice $\boldsymbol{X}$ glede na množico $\boldsymbol{\mathcal{B}}$.

Ocena napovednega modela

Implementacije napovednega modela bomo ocenjevali skozi več kriterijev, ki jih bomo s pomočjo napovedne množice $\boldsymbol{A_i}$, testne množice $\boldsymbol{B_i}$ in množic $\boldsymbol{T_i^+}$, $\boldsymbol{T_i^-}$, $\boldsymbol{F_i^+}$, $\boldsymbol{F_i^-}$ definirali v nadaljevanju.

Recimo, da je moč množice $\boldsymbol{T_i^+}$, definirane v 3.3.1, naša ocena, saj predstavlja število pravilno napovedanih kupcev za izdelek $\boldsymbol{i}$. Vendar pa je ta ocena sama po sebi slaba, saj je absolutna. Absolutne ocene lahko določene primere precenijo in druge podcenijo. Poglejmo si primer 3, kjer z oceno moči množice $\boldsymbol{T^+}$ podcenimo dejansko vrednost napovedi.

Primer 3. *Predpostavimo, da je moč množice $\boldsymbol{T_i^+}$ za nek izdelek i enaka 1. To pomeni da smo pravilno napovedali le enega kupca. Na prvi pogled se nam zdi, da je ocena zelo slaba. Vendar, če za primer velja, da je moč množice $\boldsymbol{B_i}$ prav tako 1, lahko sklepamo, da smo z napovedjo odkrili vse kupce, ki so naročili izdelek $\boldsymbol{i}$.*

Da bi oceno izboljšali, jo spremenimo v relativno glede na moč množice $\boldsymbol{B_i}$.

Definicija 3.3.5 (Prvi ocenjevalni kriterij)**.** Delež odkritih kupcev proti številu dejanskih kupcev izdelka $\boldsymbol{i}$ je enak:

$$\frac{|\boldsymbol{T_i^+}|}{|\boldsymbol{B_i}|} = \frac{|\boldsymbol{A_i} \cap \boldsymbol{B_i}|}{|\boldsymbol{B_i}|} \tag{3.1}$$

Izboljšana ocena ima sedaj razpon realnih vrednosti med 0 in 1, pri čemer 0 pomeni nič pravilno napovedanih kupcev, 1 pa, da so bili vsi dejanski kupci iz testne množice odkriti. Ta ocena je tudi naš prvi ocenjevalni kriterij, ki ga bomo uporabljali pri ocenjevanju praktičnih primerov.

Ker v oceni ne upoštevamo velikosti množice $\boldsymbol{A_i}$, lahko za množico $\boldsymbol{A_i}$ vedno vzamemo kar celo množico kupcev $\boldsymbol{\mathcal{B}}$ in prvi ocenjevalni kriterij bo vedno optimalen.

Izrek 3.1. Če velja $\boldsymbol{A_i} = \boldsymbol{\mathcal{B}}$, potem $\frac{|\boldsymbol{T_i^+}|}{|\boldsymbol{B_i}|} = 1$.

Dokaz. Velja $\boldsymbol{A_i} = \boldsymbol{\mathcal{B}}$ in $\boldsymbol{B_i} \subseteq \boldsymbol{\mathcal{B}}$. Torej dobimo:

$$\frac{|\boldsymbol{T_i^+}|}{|\boldsymbol{B_i}|} = \frac{|\boldsymbol{\mathcal{B}} \cap \boldsymbol{B_i}|}{|\boldsymbol{B_i}|} = \frac{|\boldsymbol{B_i}|}{|\boldsymbol{B_i}|} = 1.$$

□

V večini teh primerov, ko velja $|\boldsymbol{B_i}| \ll |\boldsymbol{\mathcal{B}}|$, s prvim ocenjevalnim kriterijem precenjujemo dejansko vrednost primera, saj ne upoštevamo negativnega vpliva napovedanih kupcev, ki niso del testne množice kupcev $\boldsymbol{B_i}$. Da bi odpravili to pomanjkljivost ocenjevanja, smo uvedli drugi ocenjevalni kriterij, zapisan v definiciji 3.3.6, ki omejuje velikost množice $\boldsymbol{A_i}$ glede na množico $\boldsymbol{T_i^+}$.

Definicija 3.3.6 (Drugi ocenjevalni kriterij)**.** Delež pravilno napovedanih kupcev proti vsem napovedanim kupcem za izdelek $\boldsymbol{i}$:

$$\frac{|\boldsymbol{T}_{\boldsymbol{i}}^{+}|}{|\boldsymbol{A}_{\boldsymbol{i}}|} = \frac{|\boldsymbol{A}_{\boldsymbol{i}} \cap \boldsymbol{B}_{\boldsymbol{i}}|}{|\boldsymbol{A}_{\boldsymbol{i}}|}. \tag{3.2}$$

Slednji kriterij ima razpon vrednosti med 0 in 1, pri čemer 0 pomeni, da v napovedni množici ni pravilno napovedanih kupcev, in 1, da so vsi kupci v napovedni množici tudi dejanski kupci izdelka $\boldsymbol{i}$. Ker želimo, da je čim več kupcev napovedanih pravilno, ga želimo maksimirati.

Za kriterije, ki jih definiramo, želimo, da v nobenem od primerov s svojo oceno ne precenijo ali podcenijo primera. Za prvi ocenjevalni kriterij že poznamo protiprimer, ki ga preceni. Podobno velja tudi za drugi ocenjevalni kriterij, ki ne upošteva neodkritih dejanskih kupcev. Vendar pa se prvi in drugi ocenjevalni kriterij med sabo izpopolnjujeta, saj odpravljata napake drug drugega. V nadaljevanju bomo definirali ocenjevalni kriterij, ki bo združil prvi in drugi ocenjevalni kriterij in ne bo podcenil ali precenil nobenega od robnih primerov.

Za tretji ocenjevalni kriterij želimo, da bi nekako združil prejšnja dva kriterija. Pri obeh prejšnjih kriterijih nas zanima razmerje množice pravilno napovedanih $\boldsymbol{T}_{\boldsymbol{i}}^{+}$ proti neki drugi množici. Zato bomo tudi pri novem kriteriju gledali razmerje pravilno napovedanih $\boldsymbol{T}_{\boldsymbol{i}}^{+}$ proti neki drugi množici. Za prvi ocenjevalni kriterij velja, da omejuje število napačno napovedanih kupcev $\boldsymbol{F}_{\boldsymbol{i}}^{+}$. Drugi ocenjevalni kriterij omejuje kupce, ki jih nismo odkrili, pa so v testni množici $\boldsymbol{F}_{\boldsymbol{i}}^{-}$. To pomeni, da želimo v novem ocenjevalnem kriteriju unijo omenjenih množic

$$\begin{aligned} \boldsymbol{F}_{\boldsymbol{i}}^{-} \cup \boldsymbol{F}_{\boldsymbol{i}}^{+} = \\ (\boldsymbol{A}_{\boldsymbol{i}} \setminus \boldsymbol{B}_{\boldsymbol{i}}) \cup (\boldsymbol{B}_{\boldsymbol{i}} \setminus \boldsymbol{A}_{\boldsymbol{i}}) = \\ (\boldsymbol{A}_{\boldsymbol{i}} \cup \boldsymbol{B}_{\boldsymbol{i}}) \setminus (\boldsymbol{A}_{\boldsymbol{i}} \cap \boldsymbol{B}_{\boldsymbol{i}}) = \\ (\boldsymbol{A}_{\boldsymbol{i}} \cup \boldsymbol{B}_{\boldsymbol{i}}) \setminus \boldsymbol{T}_{\boldsymbol{i}}^{+}, \end{aligned} \tag{3.3}$$

čim bolj omejiti.

Omejevanje množice v enačbi (3.3) pomeni, da želimo čim večjo množico $\boldsymbol{T_i^+}$ v primerjavi z velikostjo množice $\boldsymbol{A_i \cup B_i}$, kar lahko zapišemo z enačbo (3.4), ki je tretji ocenjevalni kriterij.

Definicija 3.3.7 (Tretji ocenjevalni kriterij)**.** Delež pravilno napovedanih kupcev za izdelek $\boldsymbol{i}$ proti številu kupcev, ki so v napovedni ali testni množici:

$$\frac{\boldsymbol{T_i^+}}{\boldsymbol{A_i \cup B_i}} = \frac{\boldsymbol{A_i \cap B_i}}{\boldsymbol{A_i \cup B_i}}. \tag{3.4}$$

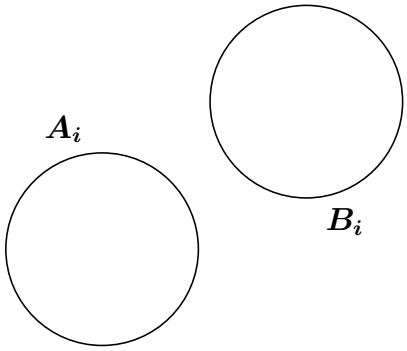
Ko se pri napovedi zmotimo in napovemo kupca, ki ni dejanski, se pojavi negativen vpliv. Primer so promocije za izdelek $\boldsymbol{i} \in \boldsymbol{\mathcal{I}}$, poslane kupcem, ki jih izdelek $\boldsymbol{i}$ ne zanima. Pri slednjem primeru negativnega vpliva lahko kot posledica pojava AdFatigue kupce odvrnemo ali izgubimo. Potemtakem lahko nepravilno napovedane kupce $\boldsymbol{F^+}$ označimo kot strošek. Na drugi strani pa imamo pravilno napovedane kupce $\boldsymbol{T^+}$, ki pa predstavljajo prihodek. Pravilno napovedani kupci so kupci, ki jih zanima izdelek $\boldsymbol{i}$, zato je zelo velika verjetnost, da se bodo pozitivno odzvali na promocijo izdelka $\boldsymbol{i}$. Ker stroški in prihodki niso v enakem razmerju, bomo mednje postavili faktor stroška $\boldsymbol{\alpha}$. Sedaj lahko definiramo dobiček, ki je hkrati tudi naš četrti ocenjevalni kriterij 3.3.8.

Definicija 3.3.8 (Četrti ocenjevalni kriterij)**.** Dobiček promocije za izdelek $\boldsymbol{i}$, ki uporablja napovedni model, je enak:

$$|\boldsymbol{T_i^+}| - \boldsymbol{\alpha}|\boldsymbol{F_i^+}|.$$

Robni primeri pri napovedovanju kupcev

V nadaljevanju bomo definirali štiri primere, ki bodo predstavljali štiri robne pogoje. Pogledali bomo, kako se pri njih obnašajo kriteriji, ki smo jih definirali. S tem bomo tudi dokazali, da tretji in četrti ocenjevalni kriterij ne precenita ali podcenita primera v nobenem od robnih primerov.



Slika 3.1: Primer $\boldsymbol{A_i} \cap \boldsymbol{B_i} = \emptyset$

Prvi robni primer Prvi robni primer se pojavi, ko napovedna množica $\boldsymbol{A_i}$ nima skupnih izdelkov s testno množico $\boldsymbol{B_i}$. Vennov diagram, ki prikazuje primer, je prikazan na sliki 3.1. Slednjo trditev oziroma predpostavko lahko zapišemo z enačbo:

$$\boldsymbol{A_i} \cap \boldsymbol{B_i} = \emptyset. \tag{3.5}$$

Iz enačbe predpostavke (3.5) in definicije 3.3.1 sledi:

$$\boldsymbol{T_i^+} = \emptyset, \tag{3.6}$$

kar pomeni, da nobenega kupca nismo napovedali pravilno. Za ta robni primer pričakujemo, da nam ga kriteriji ocenijo z najslabšo oceno.

Ko uporabimo enačbo (3.5) v enačbi (3.1) prvega ocenjevalnega kriterija, dobimo:

$$\frac{|\boldsymbol{T_i^+}|}{|\boldsymbol{B_i}|} = \frac{|\emptyset|}{|\boldsymbol{B_i}|} = \frac{0}{|\boldsymbol{B_i}|} = 0.$$

Enačba (3.2) drugega ocenjevalnega kriterija se pri upoštevanju enačbe (3.6) enači z 0 oziroma:

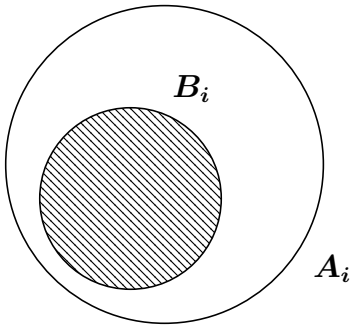
$$\frac{|\boldsymbol{T_i^+}|}{|\boldsymbol{A_i}|} = \frac{|\emptyset|}{|\boldsymbol{A_i}|} = \frac{0}{|\boldsymbol{A_i}|} = 0.$$

Uporabimo enačbo (3.5) še v enačbi (3.4) tretjega ocenjevalnega kriterija, kjer je izpeljava sledeča:

$$\frac{|\boldsymbol{T_i^+}|}{|\boldsymbol{A_i} \cup \boldsymbol{B_i}|} = \frac{|\emptyset|}{|\boldsymbol{A_i} \cup \boldsymbol{B_i}|} = \frac{0}{|\boldsymbol{A_i} \cup \boldsymbol{B_i}|} = 0.$$

Izpeljava četrtega ocenjevalnega kriterija oziroma dobička v našem primeru je:

$$|\boldsymbol{T}_i^+| - \boldsymbol{\alpha}|\boldsymbol{F}_i^+| = |\boldsymbol{A}_i \cap \boldsymbol{B}_i| - \boldsymbol{\alpha}|\boldsymbol{A}_i \setminus \boldsymbol{B}_i| = |\emptyset| - \boldsymbol{\alpha}|\boldsymbol{A}_i| = -\boldsymbol{\alpha}|\boldsymbol{A}_i|.$$



Slika 3.2: Primer $\boldsymbol{B_i} \subset \boldsymbol{A_i}$

Drugi robni primer Drugi robni primer je prikazan z diagramom na sliki 3.2. Zanj velja, da napovedna množica $\boldsymbol{A_i}$ poleg vseh kupcev iz množice $\boldsymbol{B_i}$ vsebuje še nekaj drugih nepravilno napovedanih kupcev. Tedaj velja:

$$\boldsymbol{B_i} \subset \boldsymbol{A_i}. \tag{3.7}$$

Iz predpostavke v enačbi (3.7) in definicije 3.3.1 sledi:

$$\begin{aligned} \boldsymbol{T_i^+} = \boldsymbol{A_i} \cap \boldsymbol{B_i} = \boldsymbol{B_i}, \\ \boldsymbol{A_i} \cup \boldsymbol{B_i} = \boldsymbol{A_i}. \end{aligned} \tag{3.8}$$

Interpretacija enačbe (3.8) je, da so pravilno napovedani kupci vsi kupci iz testne množice. Poglejmo, kaj se z ocenjevalnimi kriteriji zgodi pri drugem robnem primeru.

Sledi izpeljava prvega ocenjevalnega kriterija iz enačbe (3.1) in enačbe (3.8):

$$\frac{|\boldsymbol{T_i^+}|}{|\boldsymbol{B_i}|} = \frac{|\boldsymbol{B_i}|}{|\boldsymbol{B_i}|} = 1.$$

Enačba (3.2) drugega ocenjevalnega kriterija se pri upoštevanju enačbe (3.8) enači z:

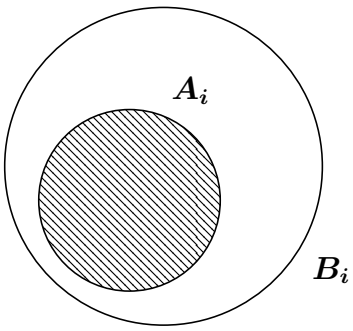
$$\frac{|\boldsymbol{T}_i^+|}{|\boldsymbol{A}_i|} = \frac{|\boldsymbol{B}_i|}{|\boldsymbol{A}_i|}.$$

Tretji ocenjevalni kriterij, izpeljan z upoštevanjem predpostavke drugega robnega primera, je enak:

$$\frac{|\boldsymbol{T}_i^+|}{|\boldsymbol{A}_i \cup \boldsymbol{B}_i|} = \frac{|\boldsymbol{B}_i|}{|\boldsymbol{A}_i \cup \boldsymbol{B}_i|} = \frac{|\boldsymbol{B}_i|}{|\boldsymbol{A}_i|}.$$

Sledi še zadnji ocenjevalni kriterij, ki predstavlja dobiček. Izpeljava zanj je enaka:

$$|\boldsymbol{T}_i^+| - \boldsymbol{\alpha}|\boldsymbol{F}_i^+| = |\boldsymbol{A}_i \cap \boldsymbol{B}_i| - \boldsymbol{\alpha}|\boldsymbol{A}_i \setminus \boldsymbol{B}_i| = |\boldsymbol{B}_i| - \boldsymbol{\alpha}|\boldsymbol{A}_i \setminus \boldsymbol{B}_i|.$$



Slika 3.3: Primer $\boldsymbol{A}_i \subset \boldsymbol{B}_i$

Tretji robni primer Tretji robni primer je podoben drugemu, le da sta množici zamenjani. Napovedna množica $\boldsymbol{A}_i$ sedaj vsebuje samo nekaj kupcev iz množice $\boldsymbol{B}_i$, kar je razvidno iz diagrama na sliki 3.3. V tem primeru je napovedna množica $\boldsymbol{A}_i$ podmnožica množice dejanskih kupcev $\boldsymbol{B}_i$ oziroma:

$$\boldsymbol{A}_i \subset \boldsymbol{B}_i. \tag{3.9}$$

Če v definiciji 3.3.1 upoštevamo enačbo (3.9) lahko izpeljemo:

$$\begin{aligned} \boldsymbol{T}_i^+ &= \boldsymbol{A}_i \cap \boldsymbol{B}_i = \boldsymbol{A}_i \\ \boldsymbol{A}_i \cup \boldsymbol{B}_i &= \boldsymbol{B}_i \end{aligned} \tag{3.10}$$

Napovedna množica $\boldsymbol{A_i}$ je zdaj enaka množici pravilno napovedanih kupcev $\boldsymbol{T_i^+}$. To pomeni, da je celotna napoved pravilna, vendar nismo odkrili vseh dejanskih kupcev iz testne množice $\boldsymbol{B_i}$

Izpeljava prvega ocenjevalnega kriterija z upoštevanjem enačbe (3.10) je naslednja:

$$\frac{|\boldsymbol{T_i^+}|}{|\boldsymbol{B_i}|} = \frac{|\boldsymbol{A_i}|}{|\boldsymbol{B_i}|}.$$

Poglejmo še drugi ocenjevalni kriterij, združen z enačbo (3.10):

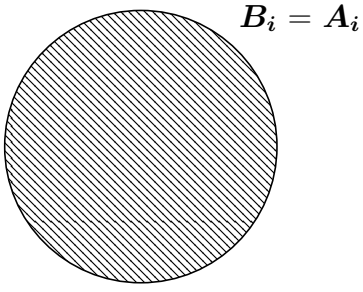
$$\frac{|\boldsymbol{T_i^+}|}{|\boldsymbol{A_i}|} = \frac{|\boldsymbol{A_i}|}{|\boldsymbol{A_i}|} = 1.$$

Izpeljava tretjega ocenjevalnega kriterija pri predpostavki tretjega robnega primera je enaka:

$$\frac{|\boldsymbol{T_i^+}|}{|\boldsymbol{A_i} \cup \boldsymbol{B_i}|} = \frac{|\boldsymbol{A_i}|}{|\boldsymbol{B_i}|}.$$

Enačba dobička oziroma četrtega ocenjevalnega kriterija, izpeljana z upoštevanjem enačbe (3.10), je:

$$|\boldsymbol{T_i^+}| - \boldsymbol{\alpha}|\boldsymbol{F_i^+}| = |\boldsymbol{A_i} \cap \boldsymbol{B_i}| - \boldsymbol{\alpha}|\boldsymbol{A_i} \setminus \boldsymbol{B_i}| = |\boldsymbol{A_i}| - \boldsymbol{\alpha}|\boldsymbol{A_i} \setminus \boldsymbol{B_i}|.$$



Slika 3.4: Primer $\boldsymbol{A_i} = \boldsymbol{B_i}$

Četrti robni primer Zadnji robni primer, prikazan na sliki 3.4, je primer, ko smo z napovedjo natančno napovedali vse kupce iz testne množice $\boldsymbol{B_i}$.

Tedaj velja, da sta množici $\boldsymbol{A_i}$ in $\boldsymbol{B_i}$ enaki, oziroma:

$$\boldsymbol{A_i} = \boldsymbol{B_i}. \tag{3.11}$$

Za definicijo o pravilno napovedanih kupcih 3.3.1 pri predpostavki (3.11) velja:

$$\boldsymbol{T_i^+} = \boldsymbol{A_i} \cap \boldsymbol{B_i} = \boldsymbol{A_i} \cup \boldsymbol{B_i} = \boldsymbol{A_i} = \boldsymbol{B_i}. \tag{3.12}$$

Množica pravilno napovedanih kupcev $\boldsymbol{T_i^+}$ ter množici $\boldsymbol{A_i}$ in $\boldsymbol{B_i}$ so popolnoma enake. To pomeni, da so vsi napovedani kupci napovedani pravilno in da so z napovedjo odkriti vsi dejanski kupci.

Za ocenjevalne kriterije želimo, da bi slednji primer ocenili najbolje. Poglejmo izpeljavo prvega ocenjevalnega kriterija, če upoštevamo enačbo (3.12):

$$\frac{|\boldsymbol{T_i^+}|}{|\boldsymbol{B_i}|} = \frac{|\boldsymbol{B_i}|}{|\boldsymbol{B_i}|} = 1.$$

Drugi ocenjevalni kriterij v četrtem robnem primeru prav tako vrne 1:

$$\frac{|\boldsymbol{T_i^+}|}{|\boldsymbol{A_i}|} = \frac{|\boldsymbol{A_i}|}{|\boldsymbol{A_i}|} = 1.$$

Tudi tretji ocenjevalni kriterij oceni četrti robni primer z najboljšo oceno:

$$\frac{|\boldsymbol{T_i^+}|}{|\boldsymbol{A_i} \cup \boldsymbol{B_i}|} = \frac{|\boldsymbol{A_i}|}{|\boldsymbol{A_i} \cup \boldsymbol{B_i}|} = \frac{|\boldsymbol{A_i}|}{|\boldsymbol{A_i}} = 1.$$

Dobiček oziroma četrti ocenjevalni kriterij je ob upoštevanjem enačbe (3.12) enak:

$$|\boldsymbol{T_i^+}| - \boldsymbol{\alpha}|\boldsymbol{F_i^+}| = |\boldsymbol{A_i} \cap \boldsymbol{B_i}| - \boldsymbol{\alpha}|\boldsymbol{A_i} \setminus \boldsymbol{B_i}| = |\boldsymbol{A_i}| - \boldsymbol{\alpha}|\emptyset| = |\boldsymbol{A_i}|.$$

Povzetek robnih primerov Poglejmo si vse štiri kriterije za vse štiri primere, zbrane v tabeli 3.1.

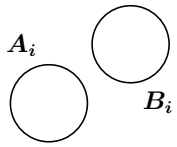
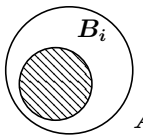
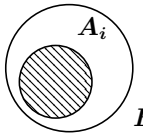
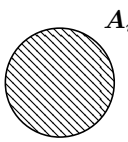
Primer	Diagram	$\frac{\lvert T_i^+\rvert}{\lvert B_i\rvert}$	$\frac{\lvert T_i^+\rvert}{\lvert A_i\rvert}$	$\frac{\lvert T_i^+\rvert}{\lvert A_i \cup B_i\rvert}$	$\lvert T_i^+\rvert - \alpha\lvert F_i^+\rvert$
$A_i \cap B_i = \emptyset$		0	0	0	$-\alpha\lvert A_i\rvert$
$B_i \subset A_i$		1	$\frac{\lvert B_i\rvert}{\lvert A_i\rvert}$	$\frac{\lvert B_i\rvert}{\lvert A_i\rvert}$	$\lvert B_i\rvert - \alpha\lvert A_i \setminus B_i\rvert$
$A_i \subset B_i$		$\frac{\lvert A_i\rvert}{\lvert B_i\rvert}$	1	$\frac{\lvert A_i\rvert}{\lvert B_i\rvert}$	$\lvert A_i\rvert - \alpha\lvert A_i \setminus B_i\rvert$
$A_i = B_i$		1	1	1	$\lvert A_i\rvert$

Tabela 3.1: Ocenjevalni kriteriji pri vseh štirih robnih primerih

Iz tabele 3.1 hitro opazimo, da so prvi trije ocenjevalni kriteriji pravilno ocenili prvi in zadnji robni primer. Vendar pa se razlike pojavijo v drugem in tretjem robnem primeru, kjer je ena od množic podmnožica drugi množici.

Pri prvem kriteriju lahko prepoznamo primer, ki smo ga opisali kot motivacijo za izpeljavo drugega ocenjevalnega kriterija. To je drugi robni primer, ko je testna množica B_i v celoti vsebovana v napovedni množici A_i. Tukaj prvi ocenjevalni kriterij s svojo oceno 1 preceni dejansko vrednost primera.

Podobno se zgodi pri drugem kriteriju in tretjem primeru, ko je napovedna množica A_i podmnožica testne množice B_i. Tedaj so vsi kupci napovedani pravilno, vendar je ostalo še nekaj neodkritih kupcev iz testne množice B_i. Tukaj drugi kriterij preceni tretji robni primer, saj ne upošteva neodkritih kupcev.

Pri tretjem kriteriju opazimo, da smo odpravili pomanjkljivosti prvega in drugega kriterija, ker vzamemo delež pravilno napovedanih proti vsem kupcem iz napovedne $\boldsymbol{A_i}$ in testne množice $\boldsymbol{B_i}$. Za tretji kriterij smo prikazali, da pri nobenem od štirih robnih primerov ne precenjuje ali podcenjuje primera.

Ocena napovednega modela nad vsemi kupci

Do sedaj smo definirali kriterije, po katerih bomo ocenjevali napovedi kupcev za izdelek $\boldsymbol{i}$. V nadaljevanju dela se ne bomo osredotočali na posamezne kupce in bomo napovedi izvajali za vse izdelke $\boldsymbol{i} \in \boldsymbol{\mathcal{I}}$. Da ne bomo prikazovali ocene vsake napovedi posebej, bomo ocene čez vse napovedi povprečili. Povprečne kriterije bomo označili z enako enačbo, kot je enačba kriterija, ki ga povprečimo, vendar z razliko izpuščenih indeksov in oznak za moč množice.

Definicija 3.3.9. Povprečje prvega ocenjevalnega kriterija preko vseh napovedi za izdelke iz množice $\boldsymbol{\mathcal{I}}$ je enako:

$$\overline{\boldsymbol{T^+/B}} = \frac{1}{|\boldsymbol{\mathcal{I}}|} \sum_{\boldsymbol{i} \in \boldsymbol{\mathcal{I}}} \frac{|\boldsymbol{T_i^+}|}{|\boldsymbol{B_i}|}.$$

Definicija 3.3.10. Povprečje drugega ocenjevalnega kriterija preko vseh napovedi za izdelke iz množice $\boldsymbol{\mathcal{I}}$ je enako:

$$\overline{\boldsymbol{T^+/A}} = \frac{1}{|\boldsymbol{\mathcal{I}}|} \sum_{\boldsymbol{i} \in \boldsymbol{\mathcal{I}}} \frac{|\boldsymbol{T_i^+}|}{|\boldsymbol{A_i}|}.$$

Definicija 3.3.11. Povprečje tretjega ocenjevalnega kriterija preko vseh napovedi za izdelke iz množice $\boldsymbol{\mathcal{I}}$ zapišemo kot:

$$\overline{\boldsymbol{T^+/(A \cup B)}} = \frac{1}{|\boldsymbol{\mathcal{I}}|} \sum_{\boldsymbol{i} \in \boldsymbol{\mathcal{I}}} \frac{|\boldsymbol{T_i^+}|}{|\boldsymbol{A_i} \cup \boldsymbol{B_i}|}.$$

Za četrti ocenjevalni kriterij ni smiselno, da ga povprečimo, ker je absoluten. Zato bomo povprečili relativen dobiček na napovedanega kupca.

Definicija 3.3.12. Povprečje dobička na napovedanega kupca preko vseh napovedi za izdelke iz množice $\boldsymbol{\mathcal{I}}$ je enako:

$$\overline{(\boldsymbol{T}^{+} - \boldsymbol{\alpha}\boldsymbol{F}^{+})/\boldsymbol{A}} = \frac{1}{|\boldsymbol{\mathcal{I}}|} \sum_{\boldsymbol{i} \in \boldsymbol{\mathcal{I}}} \frac{|\boldsymbol{T}_{\boldsymbol{i}}^{+}| - \boldsymbol{\alpha}|\boldsymbol{F}_{\boldsymbol{i}}^{+}|}{|\boldsymbol{A}_{\boldsymbol{i}}|}.$$

3.4 Napovedovanje izdelkov

Problem napovedovanja izdelkov za kupca je podoben primeru napovedovanja kupcev za izdelek, le pogled je z drugega zornega kota. Kot smo storili pri definiranju prejšnjega problema, bomo tudi tukaj začeli z definiranjem resničnostne funkcije.

Definicija resničnostne funkcije

Resničnostne funkcije ne bomo definirali posebej, saj je zelo podobna definiciji resničnostne funkcije pri napovedovanju kupcev v razdelku 3.3 poglavja 3.3. Razlika je le v interpretaciji izhoda oziroma resničnega stanja. V primeru, ko napovedujemo izdelke, bomo v izhodu resničnostne funkcije $\boldsymbol{r}(\boldsymbol{I_r})$ za nekega kupca $\boldsymbol{b}$ poiskali izdelke, za katere je ta kupec imel pozitiven odziv, in jih združili v množico $\boldsymbol{I_b}$, ki ji bomo rekli testna množica. Izdelke iz množice $\boldsymbol{I_b}$ bomo od sedaj naprej imenovali tudi dejanski izdelki.

Napovedni model

Podobno kot pri napovedovanju kupcev bomo tudi tukaj definirali napovedni model, ki napoveduje izdelke za kupca $\boldsymbol{b} \in \boldsymbol{\mathcal{B}}$ iz preteklih nakupov.

Kot vhod $\boldsymbol{I_f}$ v napovedni model, kjer napovedujemo izdelke za kupca $\boldsymbol{b}$, podamo popolnoma iste podatke, kot smo jih pri napovedovanju kupcev. To so podatki o izdelkih, ki so jih kupci kupovali v preteklosti.

Izhod oziroma napoved napovednega modela je tukaj drugačna, saj napovedujemo izdelke in ne kupcev. Ta ob napovedovanju izdelkov za kupca $\boldsymbol{b}$ tvori množico napovedanih izdelkov $\boldsymbol{C_b}$.

Definirajmo relacije med napovedno in testno množico, kot smo to storili za napovedovanje kupcev:

- pravilno napovedani izdelki ... $\boldsymbol{T_b^+ = C_b \cap I_b}$,
- nenapovedani izdelki, ki niso v testni množici ... $\boldsymbol{T_b^- = (C_b \cup I_b)^{\complement}}$,
- nepravilno napovedani izdelki ... $\boldsymbol{F_b^+ = C_b \setminus T_b^+ = C_b \setminus C_b \cap I_b = C_b \setminus I_b}$,
- neodkriti izdelki ... $\boldsymbol{F_b^- = I_b \setminus T_b^+ = I_b \setminus C_b \cap I_b = I_b \setminus C_b}$.

Ocena napovednega modela

Za oceno napovedi izdelkov za kupca $\boldsymbol{b}$ bomo definirali tri kriterije. Ti kriteriji bodo ocenjevali primer glede na testno množico $\boldsymbol{I_b}$. Izpeljava kriterijev in motivacija zanje sta enaki kot pri napovedovanju kupcev, zato ju bomo tukaj izpustili.

Definicija 3.4.1 (Prvi ocenjevalni kriterij)**.** Delež pravilno napovedanih izdelkov proti številu vseh dejanskih izdelkov iz testne množice za kupca $\boldsymbol{b}$ je enak:

$$\frac{|\boldsymbol{T_b^+}|}{|\boldsymbol{I_b}|}.$$

Definicija 3.4.2 (Drugi ocenjevalni kriterij)**.** Delež pravilno napovedanih izdelkov proti številu vseh napovedanih izdelkov za kupca $\boldsymbol{b}$ je enak:

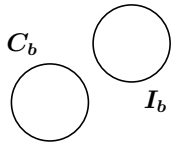
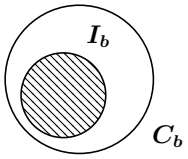
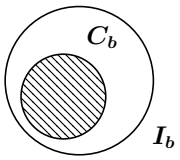
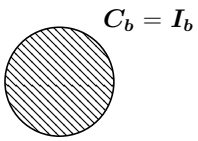
$$\frac{|\boldsymbol{T_b^+}|}{|\boldsymbol{C_b}|}.$$

Definicija 3.4.3 (Tretji ocenjevalni kriterij)**.** Delež pravilno napovedanih izdelkov za kupca $\boldsymbol{b}$ proti številu vseh izdelkov, ki so bili napovedani ali so iz testne množice, je:

$$\frac{|\boldsymbol{T_b^+}|}{|\boldsymbol{C_b \cup I_b}|}.$$

Poglejmo obnašanje kriterijev pri robnih primerih v tabeli 3.2.

Tabela 3.2: Ocenjevalni kriteriji pri vseh treh robnih primerih

Primer	Diagram	$\frac{\lvert T_b^+\rvert}{\lvert I_b\rvert}$	$\frac{\lvert T_b^+\rvert}{\lvert C_b\rvert}$	$\frac{\lvert T_b^+\rvert}{\lvert C_b \cup I_b\rvert}$
$C_b \cap I_b = \emptyset$		0	0	0
$I_b \subset C_b$		1	$\frac{\lvert I_b\rvert}{\lvert C_b\rvert}$	$\frac{\lvert I_b\rvert}{\lvert C_b\rvert}$
$C_b \subset I_b$		$\frac{\lvert C_b\rvert}{\lvert I_b\rvert}$	1	$\frac{\lvert C_b\rvert}{\lvert I_b\rvert}$
$C_b = I_b$		1	1	1

Vsi trije kriteriji so konsistentni in natančni pri prvem in zadnjem primeru. Pri ostalih dveh primerih (drugem in tretjem) pa imata prvi in drugi kriterij težave. V drugem primeru z napovedjo odkrijemo vse izdelke iz testne množice I_b, vendar pa pri napovedi zajamemo še izdelke, ki niso del testne množice. Tukaj prvi ocenjevalni kriterij preceni primer, saj ga oceni z najboljšo oceno. Težava je, ker prvi ocenjevalni kriterij ne omejuje velikosti napovedne množice C_b.

Isto se zgodi pri tretjem primeru in drugem kriteriju, kjer slednji preceni primer, ko so vsi izdelki iz napovedne množice napovedani pravilno, vendar

pa obstaja še več dejanskih izdelkov iz testne množice, ki niso bili odkriti.

Tretji ocenjevalni kriterij omenjeni napaki izniči tako, da za referenco uporablja število izdelkov iz napovedne in testne množice.

Ocena napovednega modela nad vsemi izdelki

Enako kot smo definirali povprečja kriterijev pri napovedih kupcev za izdelke, bomo definirali tudi povprečja kriterijev pri napovedih izdelkov za kupce. Oznaka povprečnega kriterija je enaka enačbi prvotnega kriterija, le z izpuščenima indeksom in oznako za moč množice.

Definicija 3.4.4 (Prvi ocenjevalni kriterij)**.** Povprečje prvega ocenjevalnega kriterija preko vseh napovedi za kupce iz množice $\boldsymbol{\mathcal{B}}$ je enako:

$$\overline{\boldsymbol{T^+/I}} = \frac{1}{|\boldsymbol{\mathcal{B}}|} \sum_{\boldsymbol{b} \in \boldsymbol{\mathcal{B}}} \frac{|\boldsymbol{T_b^+}|}{|\boldsymbol{I_b}|}.$$

Definicija 3.4.5 (Drugi ocenjevalni kriterij)**.** Povprečje drugega ocenjevalnega kriterija preko vseh napovedi za kupce iz množice $\boldsymbol{\mathcal{B}}$ je enako:

$$\overline{\boldsymbol{T^+/C}} = \frac{1}{|\boldsymbol{\mathcal{B}}|} \sum_{\boldsymbol{b} \in \boldsymbol{\mathcal{B}}} \frac{|\boldsymbol{T_b^+}|}{|\boldsymbol{C_b}|}.$$

Definicija 3.4.6 (Tretji ocenjevalni kriterij)**.** Povprečje tretjega ocenjevalnega kriterija preko vseh napovedi za kupce iz množice $\boldsymbol{\mathcal{B}}$ je enako:

$$\overline{\boldsymbol{T^+/(C \cup I)}} = \frac{1}{|\boldsymbol{\mathcal{B}}|} \sum_{\boldsymbol{b} \in \boldsymbol{\mathcal{B}}} \frac{|\boldsymbol{T_b^+}|}{|\boldsymbol{C_b \cup I_b}|}.$$

"You don't need to learn what customers say they want; you need to learn how customers behave and what they need. In other words, focus on their problem, not their suggested solution."

— Cindy Alvarez

Poglavje 4

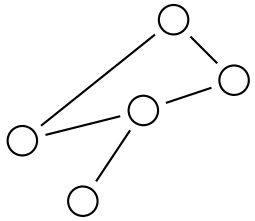
Grafovski modeli napovedovanja

V poglavju 3 smo definirali problem napovedovanja kupcev in izdelkov. V nadaljevanju bomo predstavili dva pristopa k reševanju omenjenih problemov. Napovedovanje bomo izvajali nad podatki o zgodovini nakupov. Reševanja problema napovedovanja smo se za razliko od večine drugih lotili s pristopom teorije grafov.

V prvem koraku smo kupce in izdelke povezali v neusmerjen dvodelni graf, ki nam je služil za osnovo drugega koraka. V drugem koraku smo iz dvodelnega grafa zgradili nov utežen neusmerjen graf, ki pri napovedovanju kupcev povezuje kupce, pri napovedovanju izdelkov pa izdelke. Nad grafom kupcev in izdelkov smo v tretjem koraku uporabili dva različna pristopa, s katerima smo izluščili napovedno množico kupcev pri napovedovanju kupcev in izdelkov pri napovedovanju izdelkov.

4.1 Osnovni pojmi

Preden se potopimo v podrobnosti pristopov moramo definirati osnovne pojme in nekaj znanih algoritmov, ki jih kasneje uporabljamo v predstavitvi in razvoju pristopov. Za vse grafe, ki jih bomo definirali in uporabljali,



Slika 4.1: Primer grafa

velja, da so neusmerjeni.

Graf Graf $\boldsymbol{G}$ je množica vozlišč $\boldsymbol{V}$, ki jih povezujejo povezave $\boldsymbol{E}$, ki so neurejeni pari vozlišč iz množice $\boldsymbol{V}$. Graf v osnovi zapišemo kot par množic: $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$ in ga lahko predstavimo tudi grafično, kjer so vozlišča predstavljena s točkami in povezave s črtami med točkami. Primer je prikazan na sliki 4.1. Ker je graf neusmerjen, vse povezave $(\boldsymbol{u}, \boldsymbol{v}) \in \boldsymbol{E}$ predstavljajo relacijo $(\boldsymbol{u}, \boldsymbol{v})$ kot tudi obratno relacijo $(\boldsymbol{v}, \boldsymbol{u})$ med vozlišči $\boldsymbol{u}$ in $\boldsymbol{v}$.

V nadaljevanju bomo v izrazih časovne zahtevnosti uporabljali dve oznaki. Oznaka $\boldsymbol{m}$ bo predstavljala število povezav v grafu, oznaka $\boldsymbol{n}$ pa število vozlišč, razen če bo ob navedbi časovne zahtevnosti določeno drugače.

Utežen graf Za utežen graf $\boldsymbol{G}$ velja, da ima vsaka povezava $(\boldsymbol{u}, \boldsymbol{v}) \in \boldsymbol{E}$ pripadajočo utež $\boldsymbol{w}(\boldsymbol{u}, \boldsymbol{v})$. Utež je definirana kot funkcija $\boldsymbol{w} : \boldsymbol{E} \to \mathbb{R}$, ki za vsako povezavo v grafu $\boldsymbol{G}$ vrne število, ki predstavlja utež. Utežen graf $\boldsymbol{G}$ definiramo tako, da funkcijo uteži $\boldsymbol{w}$ pripišemo paru množic ($\boldsymbol{V}$, $\boldsymbol{E}$):

$$\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E}, \boldsymbol{w}).$$

Utežen graf je običajno predstavljen z matriko sosednosti ali seznamom sosedov [19, 20]. Ker je utežen graf tudi neusmerjen, velja:

$$\forall (\boldsymbol{u}, \boldsymbol{v}) \in \boldsymbol{E} : \boldsymbol{w}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{w}(\boldsymbol{v}, \boldsymbol{u}).$$

Poln graf Graf $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$ je poln, če množica povezav $\boldsymbol{E}$ vsebuje povezave med vsemi vozlišči iz množice $\boldsymbol{V}$:

$$\forall \boldsymbol{u} \in \boldsymbol{V}, \forall \boldsymbol{v} \in \boldsymbol{V} : (\boldsymbol{u}, \boldsymbol{v}) \in \boldsymbol{E}.$$

Utežen dvodelni graf Graf, ki povezuje vozlišča iz dveh različnih množic $\boldsymbol{U}$ in $\boldsymbol{V}$ med sabo, imenujemo dvodelni graf. Za povezave $\boldsymbol{E}$ dvodelnega grafa $\boldsymbol{G}$ velja, da ne povezujejo parov znotraj množice $\boldsymbol{U}$ ali $\boldsymbol{V}$, vendar le pare elementov iz različnih množic:

$$\forall(\boldsymbol{u}, \boldsymbol{v}) \in \boldsymbol{E} : \boldsymbol{u} \in \boldsymbol{U} \wedge \boldsymbol{v} \in \boldsymbol{V}.$$

Če ima dvodelni graf s funkcijo $\boldsymbol{w}$ utežene povezave, ga imenujemo utežen dvodelni graf in definiramo kot:

$$\boldsymbol{G} = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{E}, \boldsymbol{w}).$$

Vsi pari Postopek vseh parov tvori vse kombinacije neurejenih parov elementov iz množice $\boldsymbol{U}$. Slednjega smo opisali s psevdokodo algoritma 1. Časovna zahtevnost algoritma je enaka $\boldsymbol{O}(\boldsymbol{n}^2)$.

Algoritem 1 Psevdokoda vseh parov

```
function vsiPari(U)
    if |U| < 2 then
        return ∅
    u ← prvi element množice U
    K ← ∅
    for all v ∈ U \ {u} do
        K ← K ∪ {(u, v)}
    return vsiPari(U \ {u}) ∪ K
```

Omejevanje grafa Omejevanje grafa je postopek, ki tvori podgraf uteženega neusmerjenega grafa $\boldsymbol{G}$ glede na stopnjo omejevanja $\boldsymbol{k}$. Pri omejitvi grafa $\boldsymbol{G}$ iz grafa odstranimo povezave z utežjo, manjšo od stopnje omejevanja $\boldsymbol{k}$. Postopek je opisan s psevdokodo algoritma 2. Časovna zahtevnost algoritma omejevanja grafa je $\boldsymbol{O}(\boldsymbol{m})$.

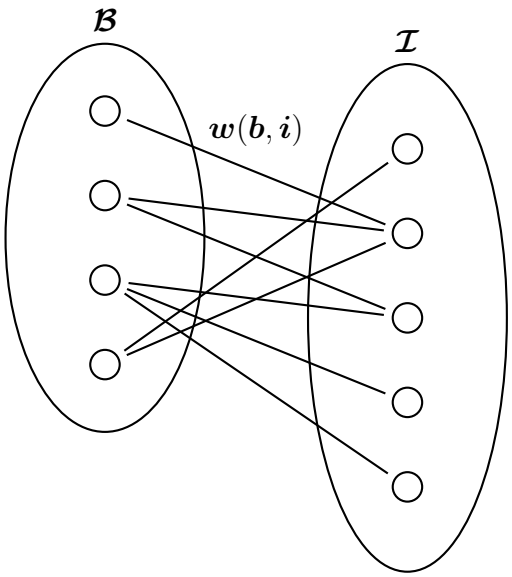
Algoritem 2 Psevdokoda omejevanja grafa $\boldsymbol{G}$

```
function omejiGraf(G, k)
    (V, E, w) ← G
    for all (u, v) ∈ E do
        if w(u, v) < k then
            E ← E \ {(u, v)}
    return (V, E, w)
```

Skupni sosedje Skupni sosedje so vsa vozlišča grafa $\boldsymbol{G}$, ki so povezana z vsemi podanimi elementi množice $\boldsymbol{U} \subseteq \boldsymbol{V}$, pri čemer je $\boldsymbol{V}$ množica vseh vozlišč grafa $\boldsymbol{G}$. Postopek iskanja skupnih sosedov v grafu je opisan z algoritmom 3, ki uporablja funkcijo sosednosti $\boldsymbol{\mathcal{N}_G(u)}$, ki vrne vse sosede elementa $\boldsymbol{u}$ v grafu $\boldsymbol{G}$. Predpostavimo, da je graf $\boldsymbol{G}$ predstavljen s seznamom sosedov ali matriko sosednosti [19, Poglavje 22]. V tem primeru je časovna zahtevnost funkcije $\boldsymbol{\mathcal{N}_G(u)}$ enaka $\boldsymbol{O(n)}$. Iz tega sledi, da ima algoritem iskanja skupnih sosedov časovno zahtevnost $\boldsymbol{O(n^2)}$, saj je velikost množice $\boldsymbol{U}$ omejena s številom $\boldsymbol{n}$ oziroma številom vozlišč v grafu $\boldsymbol{G}$.

Algoritem 3 Psevdokoda iskanja skupnih sosedov v grafu $\boldsymbol{G}$

```
function skupniSosedje(U, G)
    (V, E) ← G
    S ← V
    for all u ∈ U do
        S ← S ∩ N_G(u)
    return S
```



Slika 4.2: Dvodelni graf kupcev in izdelkov

4.2 Dvodelni graf kupcev in izdelkov

Napovedovati želimo kupce za izdelek in izdelke za kupca. Najpomembnejša člena pri našem napovedovanju sta torej kupec in izdelek. Prvi korak je, da poiščemo povezavo med kupci in izdelki ter jo predstavimo v primerni podatkovni strukturi. V našem primeru imamo na voljo podatke o preteklih naročilih, ker pa naročila že povezujejo kupce in izdelke, bodo izhodiščna točka za tvorjenje prve podatkovne strukture.

Izbrati želimo podatkovno strukturo, ki bo predstavljala povezave med kupci $\mathcal{B}$ in izdelki $\mathcal{I}$. Ko že govorimo o povezavah, lahko povezave med kupci in izdelki predstavimo kar s povezavami na grafu. Ker imamo na eni strani kupce in na drugi izdelke, je za predstavitev povezav med dvema različnima entitetama najbolj primeren utežen neusmerjen dvodelni graf (v nadaljevanju skrajšano: dvodelni graf), prikazan na sliki 4.2.

Definicija

Dvodelni graf $\boldsymbol{G_{\mathcal{BI}}}$ kupcev in izdelkov je

$$\boldsymbol{G_{\mathcal{BI}}} = (\boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{I}}, \boldsymbol{E_{\mathcal{BI}}}, \boldsymbol{w}),$$

pri čemer je $\boldsymbol{\mathcal{B}}$ množica vseh kupcev, $\boldsymbol{\mathcal{I}}$ množica vseh izdelkov in $\boldsymbol{E_{\mathcal{BI}}}$ množica s funkcijo $\boldsymbol{w}$ uteženih povezav med $\boldsymbol{\mathcal{B}}$ in $\boldsymbol{\mathcal{I}}$. Vsaka povezava med kupcem $\boldsymbol{b} \in \boldsymbol{\mathcal{B}}$ in izdelkom $\boldsymbol{i} \in \boldsymbol{\mathcal{I}}$ ima utež $\boldsymbol{w}(\boldsymbol{b}, \boldsymbol{i}) \in \mathbb{R}$, ki predstavlja, koliko izdelkov $\boldsymbol{i}$ je kupec $\boldsymbol{b}$ naročil v preteklosti. Če kupec $\boldsymbol{b}$ ni naročil izdelka $\boldsymbol{i}$ in $\boldsymbol{w}(\boldsymbol{b}, \boldsymbol{i}) = 0$, potem povezava $(\boldsymbol{b}, \boldsymbol{i})$ v grafu $\boldsymbol{G_{\mathcal{BI}}}$ ne obstaja, oziroma $(\boldsymbol{b}, \boldsymbol{i}) \notin \boldsymbol{E_{\mathcal{BI}}}$. Kupci so lahko v preteklosti naročili več različnih izdelkov, torej ima kupec lahko več povezav do izdelkov. Analogno velja tudi za izdelke, ki so lahko bili naročeni od več kupcev.

Gradnja dvodelnega grafa

Podatki, iz katerih tvorimo dvodelni graf, so naročila. Eno naročilo nosi osnovne podatke o kupcu, od kupca naročenih izdelkih, prodajalcu in količini naročenih izdelkov. Zanimajo nas naročeni izdelki, količina teh in kupec, ki jih je naročil.

Definicije oznak za naročila:

- $\boldsymbol{O}$ ... množica naročil,
- $\boldsymbol{o} \in \boldsymbol{O}$ ... posamezno naročilo,
- $\boldsymbol{q_i}$ ... število naročenih izdelkov $\boldsymbol{i}$,
- $\boldsymbol{I_o}$ ... množica parov $(\boldsymbol{i}, \boldsymbol{q_i})$, kjer je $\boldsymbol{q_i}$ količina naročenega izdelka $\boldsymbol{i}$ v naročilu $\boldsymbol{o}$.

Naročilo $\boldsymbol{o}$ je:

$$\boldsymbol{o} = (\boldsymbol{b}, \boldsymbol{I_o}),$$

kjer je $\boldsymbol{b} \in \boldsymbol{\mathcal{B}}$ kupec, ki je naročil izdelke $\boldsymbol{I_o}$.

Množica naročenih izdelkov $\boldsymbol{I_o}$ v naročilu $\boldsymbol{o}$ hrani poleg izdelkov $(\boldsymbol{i}_a, \boldsymbol{i}_b, ...)$, še naročene količine $(\boldsymbol{q}_{\boldsymbol{i}_a}, \boldsymbol{q}_{\boldsymbol{i}_b}, ...)$:

$$\boldsymbol{I_o} = \{(\boldsymbol{i}_a, \boldsymbol{q}_{\boldsymbol{i}_a}), (\boldsymbol{i}_b, \boldsymbol{q}_{\boldsymbol{i}_b}), ...\}.$$

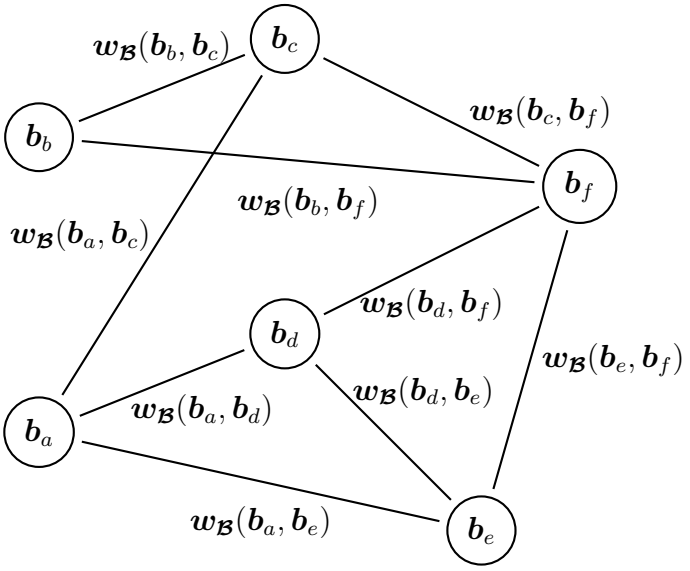
Postopek gradnje dvodelnega grafa $\boldsymbol{G_{\mathcal{BI}}}$ je sledeč. Iteriramo preko vseh naročil $\boldsymbol{O}$ in za vsako iteracijo dodamo kupca $\boldsymbol{b}$ iz naročila $\boldsymbol{o}$ v množico vseh kupcev $\boldsymbol{\mathcal{B}}$. Za vsako naročilo $\boldsymbol{o}$ iteriramo preko vseh izdelkov $\boldsymbol{I_o}$ v naročilu $\boldsymbol{o}$ in dodamo izdelek $\boldsymbol{i}$ v množico vseh izdelkov $\boldsymbol{\mathcal{I}}$. Če povezava med kupcem $\boldsymbol{b}$ in izdelkom $\boldsymbol{i}$ še ne obstaja, jo dodamo v množico povezav $\boldsymbol{E_{\mathcal{BI}}}$ in shranimo utež $\boldsymbol{w}(\boldsymbol{b}, \boldsymbol{i})$, enako količini $\boldsymbol{q_i}$. Če povezava že obstaja, utež $\boldsymbol{w}(\boldsymbol{b}, \boldsymbol{i})$ povečamo za količino $\boldsymbol{q_i}$.

Algoritem 4 Psevdokoda gradnje dvodelnega grafa $\boldsymbol{G_{\mathcal{BI}}}$

```
B ← ∅
I ← ∅
E_BI ← ∅
for all (b, I_o) ∈ O do
    B ← B ∪ {b}
    for all (i, q_i) ∈ I_o do
        I ← I ∪ {i}
        if (b, i) ∉ E_BI then
            E_BI ← E_BI ∪ {(b, i)}
            w(b, i) ← q_i
        else
            w(b, i) ← w(b, i) + q_i
G_BI ← (B, I, E_BI, w)
```

Časovna zahtevnost algoritma 4 za gradnjo dvodelnega grafa je $\boldsymbol{O}(\boldsymbol{n})$, pri čemer je $\boldsymbol{n}$ število vseh naročenih izdelkov v vseh naročilih.

Dvodelni graf kupcev in izdelkov je primeren za iskanje kupcev, ki so kupili določen izdelek $\boldsymbol{i}$, ali pa za iskanje izdelkov, ki jih je kupil poljuben kupec $\boldsymbol{b}$. Vendar pri iskanju podobnih izdelkov ali kupcev glede na zgodovino



Slika 4.3: Graf kupcev

nakupov dvodelni graf ni najboljši. Prav zato smo se odločili, da bomo zgradili dva dodatna grafa: graf kupcev in graf izdelkov.

4.3 Graf kupcev

Graf kupcev bo pomagal pri napovedovanju kupcev, kjer nas bo zanimala podobnost kupcev. Graf kupcev je neusmerjen utežen graf:

$$\boldsymbol{G_{\mathcal{B}}} = (\boldsymbol{\mathcal{B}}, \boldsymbol{E_{\mathcal{B}}}, \boldsymbol{w_{\mathcal{B}}}),$$

kjer so vozlišča vsi kupci $\boldsymbol{\mathcal{B}}$, $\boldsymbol{E_{\mathcal{B}}}$ je množica povezav med kupci in funkcija $\boldsymbol{w_{\mathcal{B}}}$ podaja utež med dvema kupcema. Primer grafa kupcev je prikazan na sliki 4.3.

Utež $\boldsymbol{w_{\mathcal{B}}}(\boldsymbol{b_u}, \boldsymbol{b_v})$ predstavlja podobnost kupcev $\boldsymbol{b_u}$ in $\boldsymbol{b_v}$. Podobnost ali utež je v osnovi število skupnih izdelkov, ki sta jih v preteklosti naročala oba kupca. Kljub temu lahko utež prilagodimo z dodatnimi parametri. Ena družina parametrov so lastnosti kupcev, na primer geografska bližina med kupcema ali razlika starosti med kupcema. Druga družina parametrov so

lastnosti izdelkov, ki sta jih kupca naročila. Če na primer oba kupca veliko kupujeta izdelke iz iste kategorije, potem je lahko njuna podobnost oziroma utež povezave med njima večja.

Gradnja grafa kupcev

Za gradnjo grafa kupcev potrebujemo množico vseh kupcev $\mathcal{B}$, za vsakega kupca $b \in \mathcal{B}$ pa moramo tudi vedeti, katere izdelke je naročal v preteklosti. Slednje lahko dobimo iz dvodelnega grafa $G_{\mathcal{BI}}$ tako, da za kupca b v grafu poiščemo vse sosede.

Z gradnjo grafa kupcev $G_{\mathcal{B}}$ začnemo s tvorjenjem množice povezav $E_{\mathcal{B}}$ med vsemi kupci množice $\mathcal{B}$. Nato za vsako povezavo $e \in E_{\mathcal{B}}$ izvedemo naslednji postopek. Za par kupcev (b_u, b_v) v povezavi e poiščemo skupne sosede na dvodelnem grafu $G_{\mathcal{BI}}$, s čimer pridobimo skupne izdelke kupcev b_u in b_v. Sedaj uporabimo izbrano funkcijo uteži, ki iz para kupcev b_u, b_v in njunih skupnih izdelkov izračuna utež $w_{\mathcal{B}}(b_u, b_v)$. Po omenjenem postopku imamo utežen poln graf kupcev $G_{\mathcal{B}}$.

Algoritem 5 Psevdokoda gradnje grafa $G_{\mathcal{B}}$

```
E_B ← VSIPARI(B)
for all (b_u, b_v) ∈ E_B do
    izdelki ← SKUPNISOSEDJE({b_u, b_v}, G_BI)
    w_B(b_u, b_v) ← IZRACUNAJUTEZ(b_u, b_v, izdelki)
G_B ← (B, E_B, w)
```

Poln graf kupcev $G_{\mathcal{B}}$ vsebuje povezave med vsemi kupci, kar namiguje na to, da so si vsi kupci med sabo podobni, vendar si niso. Zato imamo na povezavah uteži, ki nam povedo, kako močna je povezava oziroma podobnost med kupcema. Smiselno je, da preden graf uporabimo, povezave z dovolj nizko utežjo odstranimo. Omenjeno akcijo dosežemo s postopkom omejevanjem grafa, ki smo ga definirali z algoritmom 2. Po omejevanju polnega grafa $G_{\mathcal{B}}$ dobimo podgraf, ki ni več poln in ima vse povezave močnejše

ali enake stopnji omejevanja $\boldsymbol{k}$. Z variiranjem stopnje omejevanja $\boldsymbol{k}$ dobimo različne podgrafe, ki jih kasneje uporabimo v pristopih za napovedovanje. Ker potemtakem stopnja omejevanja $\boldsymbol{k}$ vpliva na napovedi, je $\boldsymbol{k} \in \mathbb{N}$ vhodni parameter.

Funkcije uteži

Nosilci najbolj pomembne informacije o podobnosti kupcev so uteži na povezavah v grafu $\boldsymbol{G_{\mathcal{B}}}$. Pravzaprav so uteži tiste, ki preko omejevanja grafa vplivajo na obliko grafa. Pomembno je, da pri računanju uteži za povezavo poskušamo upoštevati dejavnike, ki najbolj vplivajo na dejstvo, da bi povezana kupca naročila isti izdelek.

Za računanje uteži povezave med kupcema $\boldsymbol{b_u}$ in $\boldsymbol{b_v}$ imamo na voljo njune skupne izdelke, ki sta jih kupovala v preteklosti. S pomočjo dodatnih informacij lahko tvorimo več funkcij uteži. Sledijo opisi nekaj primerov takih funkcij.

Utež 1 (Število skupnih izdelkov) je najbolj enostavna funkcija. Za utež uporabi število skupnih izdelkov kupcev $\boldsymbol{b_u}$ in $\boldsymbol{b_v}$, kar lahko zapišemo tudi kot: $|\boldsymbol{\mathcal{N}_{\mathcal{BI}}}(\boldsymbol{b_u}) \cap \boldsymbol{\mathcal{N}_{\mathcal{BI}}}(\boldsymbol{b_v})|$ ali $|\textsc{skupniSosedje}(\{\boldsymbol{b_u}, \boldsymbol{b_v}\}, \boldsymbol{G_{\mathcal{BI}}})|$.

Utež 2 (Število skupnih izdelkov z naročenimi količinami) je funkcija, zelo podobna funkciji za utež 1, le da upošteva še količino naročenih izdelkov, ki sta jih kupca $\boldsymbol{b_u}$ in $\boldsymbol{b_v}$ naročala v preteklosti. Omenjeno količino naročenih izdelkov lahko dobimo iz dvodelnega grafa $\boldsymbol{G_{\mathcal{BI}}}$, kjer količino naročenih izdelkov predstavljajo uteži povezav med kupci in izdelki. Psevdokoda funkcije je predstavljena v algoritmu 6.

Utež 3 (Skupne kategorije) je funkcija, ki uporablja dodatno informacijo o izdelkih. Za vse skupne izdelke pridobi njihove kategorije in jih prešteje. Število izdelkov deli s številom preštetih kategorij, kar vrne povprečno število skupnih izdelkov v kategorijah. Dobljeno povprečje je utež 3. Postopek je opisan tudi z algoritmom 7.

Algoritem 6 Psevdokoda uteži 2

```
function IzracunajUtez2(b_u, b_v, izdelki)
    skupnaUtez ← 0
    for all i ∈ izdelki do
        skupnaUtez ← skupnaUtez + w(b_u, i)
        skupnaUtez ← skupnaUtez + w(b_v, i)
    return skupnaUtez
```

Algoritem 7 Psevdokoda uteži 3

```
function IzracunajUtez3(b_u, b_v, izdelki)
    kategorije ← ∅
    for all i ∈ izdelki do
        c ← kategorija izdelka i
        if c ∉ kategorije then
            kategorije ← kategorije ∪ c
    return |izdelki| / |kategorije|
```

Utež 4 (Geografska razdalja med kupcema) je primer funkcije, ki za utež vzame geografsko razdaljo med kupcema v povezavi. Utež 4 je predstavljena zgolj kot primer in je nismo implementirali, saj nimamo natančnih podatkov o lokaciji kupcev.

Opisali smo 4 različne funkcije za računanje uteži, vendar jih obstaja veliko več. Omejuje jih le količina dodatnih informacij, ki jih lahko pridobimo o kupcih ali izdelkih. Uporabljamo lahko tudi več uteži hkrati, tako da jih kombiniramo med sabo. Ko kombiniramo dve ali več uteži med sabo, v bistvu tvorimo novo utež, ki združi funkcije uteži z aritmetično operacijo ali algoritmom.

Primer je funkcija uteženih uteži, ki združi več funkcij tako, da izhod vsake funkcije uteži z določenim faktorjem. Funkcija uteženih uteži, ki združuje utež 1 in utež 2 v razmerju 4:6, je funkcija, definirana z algoritmom 8.

Algoritem 8 Psevdokoda združenih uteži 1 in uteži 2

```
function IzracunajUtez1Utez2(b_a, b_b, izdelki)
    return IzracunajUtez1(b_a, b_b, izdelki) * 0,4
     + IzracunajUtez2(b_a, b_b, izdelki) * 0,6
```

Različne funkcije uteži vplivajo na uteži v grafu in posredno preko omejevanja grafa tudi na podgraf, ki ga bomo v nadaljevanju uporabljali za napovedovanje kupcev. Zato izbira funkcije uteži velja za vhodni parameter.

4.4 Graf izdelkov

Naslednja struktura, ki jo bomo definirali, je neusmerjen graf izdelkov (Slika 4.4):

$$\boldsymbol{G_{\mathcal{I}}} = (\boldsymbol{\mathcal{I}}, \boldsymbol{E_{\mathcal{I}}}, \boldsymbol{w_{\mathcal{I}}}).$$

Vozlišča v grafu so izdelki iz množice vseh izdelkov $\boldsymbol{\mathcal{I}}$ in $\boldsymbol{E_{\mathcal{I}}}$ je množica povezav med izdelki, ki so utežene s funkcijo:

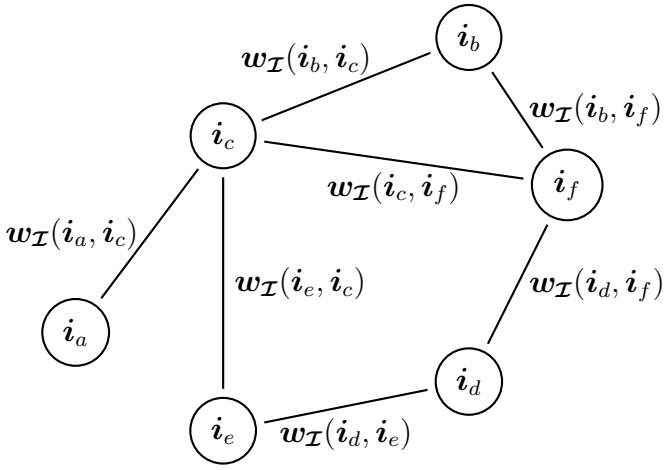
$$\boldsymbol{w_{\mathcal{I}}} : \boldsymbol{E_{\mathcal{I}}} \rightarrow \mathbb{R}.$$

Uteži na povezavah $\boldsymbol{E_{\mathcal{I}}}$ grafa $\boldsymbol{G_{\mathcal{I}}}$ predstavljajo podobnost izdelkov. Izdelek $\boldsymbol{i_u}$ in $\boldsymbol{i_v}$ sta si podobna, če med njima obstaja povezava $(\boldsymbol{i_u}, \boldsymbol{i_v})$. Večja kot je utež $\boldsymbol{w_{\mathcal{I}}}(\boldsymbol{i_u}, \boldsymbol{i_v})$ povezave, večja je podobnost izdelkov $\boldsymbol{i_u}$ in $\boldsymbol{i_v}$. Podobno kot pri utežeh kupcev lahko tudi tukaj za utež upoštevamo različne parametre, kot so dodatne informacije o izdelkih (cena, kategorija ...), kupcih ali naročilih. V osnovi sta dva izdelka podobna, če obstaja kupec, ki je v preteklosti naročil oba izdelka.

Gradnja grafa izdelkov

Postopek gradnje grafa izdelkov je podoben postopku gradnje grafa kupcev. Vir je prav tako isti dvodelni graf kupcev in izdelkov $\boldsymbol{G_{\mathcal{BI}}}$.

Prvi korak pri gradnji grafa izdelkov $\boldsymbol{G_{\mathcal{I}}}$ je tvorba množice povezav med vsemi izdelki $\boldsymbol{\mathcal{I}}$. V naslednjem koraku iteriramo preko vseh povezav $(\boldsymbol{i_u}, \boldsymbol{i_v}) \in$



Slika 4.4: Graf izdelkov

$\boldsymbol{E_{\mathcal{I}}}$, ki smo jih tvorili, in v vsaki iteraciji iz dvodelnega grafa $\boldsymbol{G_{\mathcal{BI}}}$ poiščemo skupne sosede izdelkov $\boldsymbol{i_u}$ in $\boldsymbol{i_v}$. Ti sosedje so kupci, ki so v preteklosti naročili oba izdelka. S pomočjo izbrane funkcije uteži nato izračunamo utež $\boldsymbol{w_{\mathcal{I}}(i_u, i_v)}$ za povezavo $(\boldsymbol{i_u}, \boldsymbol{i_v}) \in \boldsymbol{E_{\mathcal{I}}}$. Z opisanim postopkom ustvarimo poln utežen neusmerjen graf izdelkov.

Algoritem 9 Psevdokoda gradnje grafa $\boldsymbol{G_{\mathcal{I}}}$

$\boldsymbol{E_{\mathcal{I}}} \leftarrow$ VSIPARI($\boldsymbol{\mathcal{I}}$)
for all $(\boldsymbol{i_u}, \boldsymbol{i_v}) \in \boldsymbol{E_{\mathcal{I}}}$ **do**
    $kupci \leftarrow$ SKUPNISOSEDJE($\{\boldsymbol{i_u}, \boldsymbol{i_v}\}, \boldsymbol{G_{\mathcal{BI}}}$)
    $\boldsymbol{w_{\mathcal{I}}(i_u, i_v)} \leftarrow$ IZRACUNAJUTEZ($\boldsymbol{i_u}, \boldsymbol{i_v}, kupci$)
$\boldsymbol{G_{\mathcal{I}}} \leftarrow (\boldsymbol{\mathcal{I}}, \boldsymbol{E_{\mathcal{I}}})$

Kot pri gradnji grafa kupcev je smiselno graf izdelkov $\boldsymbol{G_{\mathcal{I}}}$ omejiti, preden ga uporabimo v nadaljevanju. Omejimo ga s postopkom omejevanja grafov, opisanim v algoritmu 2. Po omejevanju polnega grafa izdelkov $\boldsymbol{G_{\mathcal{I}}}$ dobimo podgraf, ki ima vse povezave močnejše ali enake stopnji omejevanja $\boldsymbol{k}$.

Funkcije uteži

V algoritmu 9 smo pri gradnji grafa kupcev $\boldsymbol{G_{\mathcal{I}}}$ klicali funkcijo IzracunajUtez, ki pa ima lahko več implementacij. Poglejmo si nekaj primerov funkcij za računanje uteži pri grafu kupcev $\boldsymbol{G_{\mathcal{I}}}$.

Utež 1 (Število skupnih kupcev) je najbolj enostavna funkcija, ki za utež uporabi število skupnih kupcev izdelkov $\boldsymbol{i_u}$ in $\boldsymbol{i_v}$, oziroma $|\boldsymbol{\mathcal{N}_{\mathcal{BI}}}(\boldsymbol{i_u}) \cap \boldsymbol{\mathcal{N}_{\mathcal{BI}}}(\boldsymbol{i_v})|$.

Utež 2 (Število skupnih kupcev z naročenimi količinami) je funkcija, zapisana z algoritmom 10, ki poleg števila kupcev izdelkov $\boldsymbol{i_u}$, $\boldsymbol{i_v}$ iz naročil upošteva še količine izdelkov $\boldsymbol{i_u}$ in $\boldsymbol{i_v}$, ki so jih kupci naročali.

Algoritem 10 Psevdokoda uteži 2

```
function IzracunajUtez2(i_u, i_v, kupci)
    skupnaUtez ← 0
    for all b ∈ kupci do
        skupnaUtez ← skupnaUtez + w(b, i_u)
        skupnaUtez ← skupnaUtez + w(b, i_v)
    return skupnaUtez
```

Utež 3 (Skupne kategorije) je zelo enostavna funkcija, različna od funkcije uteži 3 pri grafu kupcev. Funkcija vrne 1, če sta izdelka $\boldsymbol{i_u}$ in $\boldsymbol{i_v}$ iz iste kategorije, drugače pa vrne 0.

Utež 4 (Geografska razdalja med kupci) je primer funkcije, ki za utež vzame povprečno geografsko razdaljo med skupnimi kupci izdelkov $\boldsymbol{i_u}$ in $\boldsymbol{i_v}$. Funkcija uteži 4 je zgolj primer in je zaradi odsotnosti geografskih podatkov o kupcih nismo implementirali.

Na enak način, kot smo to opisali v razdelku 4.3 pri funkcijah uteži kupcev, lahko tudi tukaj kombiniramo različne uteži med sabo.

4.5 Napovedovanje kupcev

Pri napovedovanju kupcev iščemo kupce $\boldsymbol{A_i}$, za katere napovedujemo, da bodo v prihodnosti naročili izdelek $\boldsymbol{i}$. Iz dvodelnega grafa $\boldsymbol{G_{\mathcal{BI}}}$ lahko za izdelek $\boldsymbol{i}$ pridobimo kupce $\boldsymbol{B_i}$, ki so v preteklosti kupovali izdelek $\boldsymbol{i}$. Z napovedjo želimo poiskati kupce, ki so čim bolj podobni kupcem iz množice $\boldsymbol{B_i}$, saj je velika verjetnost, da bo podobnim kupcem izdelek $\boldsymbol{i}$ zanimiv. Podobnost vseh kupcev $\boldsymbol{\mathcal{B}}$ hranimo v grafu $\boldsymbol{G_{\mathcal{B}}}$ kot uteži na povezavah med kupci. Uteži grafa $\boldsymbol{G_{\mathcal{B}}}$ so odvisne od funkcije uteži, ki smo jo uporabili pri gradnji grafa $\boldsymbol{G_{\mathcal{B}}}$, zaradi česar bomo za lažjo razlago v tem razdelku predpostavili, da za gradnjo grafa kupcev uporabljamo utež 1 oziroma število skupnih izdelkov. Tedaj velja, da sta si dva kupca podobna, če imata podobno preteklost naročil.

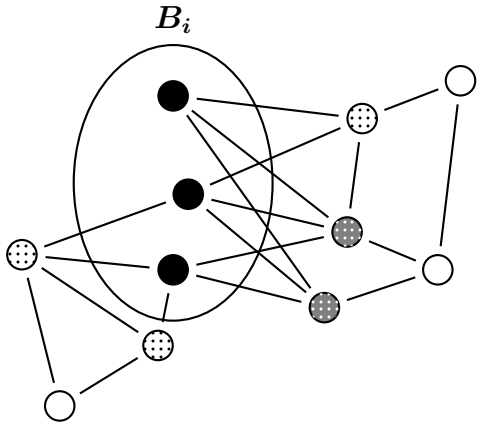
Podobne kupce iz množice $\boldsymbol{B_i}$ lahko poiščemo tako, da zanje poiščemo sosede v grafu $\boldsymbol{G_{\mathcal{B}}}$. Tukaj se pojavita dve možnosti oziroma dva pristopa. Prvi pristop je, da vzamemo skupne sosede preteklih kupcev $\boldsymbol{B_i}$, medtem ko pri drugem pristopu vzamemo vse sosede preteklih kupcev $\boldsymbol{B_i}$ izdelka $\boldsymbol{i}$.

Pristop 1: Skupni sosedje

Poglejmo si celoten postopek napovedovanja kupcev za izdelek $\boldsymbol{i}$ s prvim pristopom. Kot smo že omenili, najprej za izdelek $\boldsymbol{i}$ poiščemo pretekle dejanske kupce $\boldsymbol{B_i}$. To storimo tako, da iz dvodelnega grafa $\boldsymbol{G_{\mathcal{BI}}}$ vzamemo sosede izdelka $\boldsymbol{i}$. V naslednjem koraku lahko uporabimo funkcijo skupnih sosedov, ki smo jo definirali v razdelku 4.1 tega poglavja, vendar pa bomo za lažjo predstavo in primerjavo pristopov iskanje skupnih sosedov v algoritmu prvega pristopa 11 implementirali ponovno. Za vsakega preteklega kupca $\boldsymbol{b}$ iz množice $\boldsymbol{B_i}$ poiščemo sosede v grafu $\boldsymbol{G_{\mathcal{B}}}$. Nato naredimo presek množic sosedov vsakega kupca $\boldsymbol{b} \in \boldsymbol{B_i}$. Dobljena množica je napovedna množica $\boldsymbol{A_i}$, ki vsebuje kupce, za katere napovedujemo, da bodo v prihodnosti naročili izdelek $\boldsymbol{i}$. Postopek je opisan tudi s psevdokodo v algoritmu 11.

Algoritem 11 Prvi pristop pri napovedovanju kupcev

$\boldsymbol{B_i} \leftarrow \boldsymbol{\mathcal{N}_{\mathcal{BI}}}(\boldsymbol{i})$
$\boldsymbol{A_i} \leftarrow \boldsymbol{\mathcal{B}}$
for all $\boldsymbol{b} \in \boldsymbol{B_i}$ **do**
    $\boldsymbol{A_i} \leftarrow \boldsymbol{A_i} \cap \boldsymbol{\mathcal{N}_{\mathcal{B}}}(\boldsymbol{b})$



Slika 4.5: Skupni in vsi sosedje množice $\boldsymbol{B_i}$

Pristop 2: Vsi sosedje

Opišimo še postopek drugega pristopa za napovedovanje kupcev izdelka $\boldsymbol{i}$. Za izdelek $\boldsymbol{i}$ iz grafa $\boldsymbol{G_{\mathcal{BI}}}$ poiščemo sosede, ki tvorijo množico preteklih dejanskih kupcev $\boldsymbol{B_i}$ izdelka $\boldsymbol{i}$. Za vsakega kupca $\boldsymbol{b}$ iz množice $\boldsymbol{i}$ poiščemo sosede ter dobljene množice z unijo združimo v napovedno množico $\boldsymbol{A_i}$. Tukaj je napovedna množica večja kot pri prvem pristopu, saj zajema več sosedov in ne samo skupnih. To je tudi razvidno iz grafa na sliki 4.5, kjer so črna vozlišča elementi množice $\boldsymbol{B_i}$, sivi vozlišči skupna soseda ter vozlišča s pikami (vključno s skupnima sosedoma) vsi sosedje preteklih kupcev $\boldsymbol{B_i}$. Drugi pristop smo opisali z algoritmom 12.

Algoritem 12 Drugi pristop pri napovedovanju kupcev

$\boldsymbol{B_i} \leftarrow \boldsymbol{\mathcal{N}_{\mathcal{BI}}(i)}$
$\boldsymbol{A_i} \leftarrow \boldsymbol{\emptyset}$
for all $\boldsymbol{b} \in \boldsymbol{B_i}$ **do**
    $\boldsymbol{A_i} \leftarrow \boldsymbol{A_i} \cup \boldsymbol{\mathcal{N}_{\mathcal{B}}(b)}$

4.6 Napovedovanje izdelkov

Za napovedovanje izdelkov velja, da za kupca $\boldsymbol{b}$ iščemo najprimernejše izdelke. Pristopa, ki ju bomo predstavili v nadaljevanju, sta enaka kot pri napovedovanju kupcev za izdelek, vendar ju zaradi drugačne logične predstavitve in različnih uporabljenih podatkovnih struktur opisujemo posebej.

Pristop 1: Skupni sosedje

V prvem koraku želimo za kupca $\boldsymbol{b}$ pridobiti izdelke $\boldsymbol{I_b}$, ki jih je kupec $\boldsymbol{b}$ kupoval v preteklosti. To storimo tako, da v dvodelnem grafu $\boldsymbol{G_{\mathcal{BI}}}$ poiščemo sosede kupca $\boldsymbol{b}$. V naslednjem koraku uporabimo graf $\boldsymbol{G_{\mathcal{I}}}$, iz katerega pridobimo skupne sosede izdelkov v množici $\boldsymbol{I_b}$. Dobljena množica izdelkov je napovedna množica $\boldsymbol{C_b}$. Podroben postopek je prikazan s psevdokodo v algoritmu 13.

Algoritem 13 Prvi pristop, pri napovedovanju izdelkov

$\boldsymbol{I} \leftarrow \boldsymbol{\mathcal{N}_{\mathcal{BI}}(b)}$
$\boldsymbol{C_b} \leftarrow \boldsymbol{\mathcal{I}}$
for all $\boldsymbol{i} \in \boldsymbol{B_i}$ **do**
    $\boldsymbol{C_i} \leftarrow \boldsymbol{C_i} \cap \boldsymbol{\mathcal{N}_{\mathcal{I}}(i)}$

Pristop 2: Vsi sosedje

Celoten postopek pristopa vseh sosedov, opisan tudi z algoritmom 14, se prav tako začne s tvorjenjem množice izdelkov $\boldsymbol{I_b}$, ki jih je kupec $\boldsymbol{b}$ kupoval

v preteklosti. V drugem koraku za vse kupce iz množice $\boldsymbol{I_b}$ s pomočjo grafa $\boldsymbol{G_{\mathcal{I}}}$ poiščemo sosede ter jih z unijo združimo v napovedno množico $\boldsymbol{C_b}$.

Algoritem 14 Drugi pristop, pri napovedovanju izdelkov

$\boldsymbol{I} \leftarrow \boldsymbol{\mathcal{N}_{\mathcal{BI}}}(\boldsymbol{b})$
$\boldsymbol{C_b} \leftarrow \emptyset$
for all $\boldsymbol{i} \in \boldsymbol{I_b}$ **do**
    $\boldsymbol{C_b} \leftarrow \boldsymbol{C_b} \cup \boldsymbol{\mathcal{N}_{\mathcal{I}}}(\boldsymbol{i})$

4.7 Hibridna metoda napovedovanja kupcev

Metodi napovedovanja kupcev in napovedovanja izdelkov lahko združimo v hibridno metodo napovedovanja kupcev. V postopku hibridne metode najprej napovemo kupce za izdelek $\boldsymbol{i}$, v drugem koraku pa za vse kupce napovemo izdelke. Tretji korak je združitev napovedi kupcev za izdelek $\boldsymbol{i}$ z napovedanimi izdelki za vse kupce $\boldsymbol{\mathcal{B}}$. Množico s hibridno metodo napovedanih kupcev za izdelek $\boldsymbol{i}$ bomo označili z $\boldsymbol{A_i^*}$. Postopek združitve napovedi lahko izvedemo na dva načina.

Prvi način je pristop unije, kjer upoštevamo napovedi kupcev za izdelek $\boldsymbol{i}$ skupaj z napovedmi izdelkov. Za izdelek $\boldsymbol{i}$ in napovedanega kupca $\boldsymbol{b} \in \boldsymbol{A_i^*}$ velja, da je kupec $\boldsymbol{b}$ del množice napovedanih kupcev $\boldsymbol{A_i}$ za izdelek $\boldsymbol{i}$ ali pa je izdelek $\boldsymbol{i}$ del množice napovedanih izdelkov $\boldsymbol{C_b}$ za kupca $\boldsymbol{b}$, oziroma:

$$\forall \boldsymbol{b} \in \boldsymbol{A_i^*} : \boldsymbol{b} \in \boldsymbol{A_i} \vee \boldsymbol{i} \in \boldsymbol{C_b}.$$

Drugi način je pristop preseka, kjer se metodi napovedovanja potrjujeta. V tem primeru mora napoved kupca $\boldsymbol{b}$ za izdelek $\boldsymbol{i}$ pri napovedovanju kupcev potrditi napoved izdelka $\boldsymbol{i}$ za kupca $\boldsymbol{b}$ pri napovedovanju izdelkov. V tem primeru za izdelek $\boldsymbol{i}$ in napovedanega kupca $\boldsymbol{b} \in \boldsymbol{A_i^*}$ velja, da je kupec $\boldsymbol{b}$ del množice napovedanih kupcev kot tudi, da je izdelek $\boldsymbol{i}$ del napovedne množice $\boldsymbol{C_b}$ za kupca $\boldsymbol{b}$. Analogen zapis pristopa preseka je:

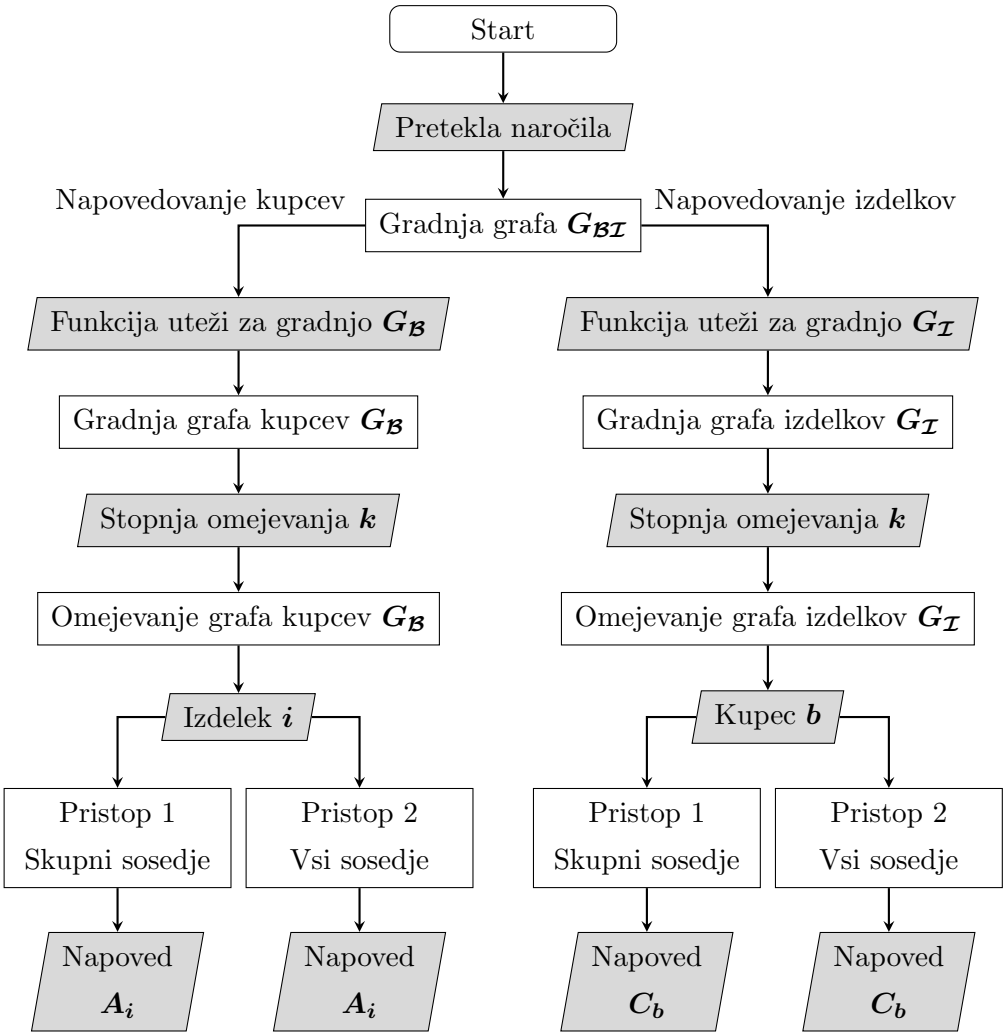
$$\forall \boldsymbol{b} \in \boldsymbol{A_i^*} : \boldsymbol{b} \in \boldsymbol{A_i} \wedge \boldsymbol{i} \in \boldsymbol{C_b}.$$

4.8 Povzetek pristopov

Za napovedovanje izdelkov smo definirali dva, za napovedovanje kupcev pa tri pristope. V prvem pristopu obeh napovedi je za tvorjenje napovedne množice uporabljen algoritem skupnih sosedov, v drugem pristopu pa algoritem vseh sosedov. V tretjem pristopu napovedovanja kupcev smo poleg pristopa, ki se je po testih izkazal za boljšega, uporabili še napovedi izdelkov. Vhod pri obeh prvih pristopih sta dva grafa. Prvi graf je dvodelni graf $\boldsymbol{G_{\mathcal{BI}}}$, drugi graf pa je graf kupcev $\boldsymbol{G_{\mathcal{B}}}$, ko napovedujemo kupce, ali pa graf izdelkov $\boldsymbol{G_{\mathcal{I}}}$, ko napovedujemo izdelke. Pri gradnji grafa kupcev $\boldsymbol{G_{\mathcal{BI}}}$ kot tudi pri gradnji grafa izdelkov $\boldsymbol{G_{\mathcal{I}}}$ uporabljamo funkcijo uteži, ki je prvi vhodni parameter. Preden nad grafom kupcev $\boldsymbol{G_{\mathcal{B}}}$ ali grafom izdelkov $\boldsymbol{G_{\mathcal{I}}}$ uporabimo algoritem napovedovanja, graf še omejimo s stopnjo omejevanja $\boldsymbol{k}$, ki je drugi vhodni parameter.

Na sliki 4.6 je prikazan diagram poteka za napovedovanje kupcev in izdelkov, ki povzame rešitev, ki smo jo razvili. Iz diagrama je razvidno, da ob spreminjanju vhodnih parametrov, kot so funkcija uteži, stopnja omejevanja ali izbira izdelka $\boldsymbol{i}$ oziroma kupca $\boldsymbol{b}$, ni potrebno ponovno izvajati celotnega postopka.

Vsak pristop bomo ovrednotili z ocenjevalnimi kriteriji, ki smo jih definirali v poglavju 3, in ocene skupaj z rezultati in s primerjavami predstavili v poglavju 6.



Slika 4.6: Diagram poteka naše rešitve

Poglavje 5

Priprava podatkov

V nadaljevanju bomo povzeli postopek zbiranja, procesiranja in čiščenja podatkov, ki smo jih kasneje uporabili pri testiranju rešitve, razvite v sklopu magistrske naloge. Po zbiranju in čiščenju bomo podatke združili s podatki iz dodatnega vira.

Zbiranje, procesiranje, čiščenje in združevanje podatkov so koraki, ki se jih uporablja v podatkovnih analizah oziroma raziskavah [21, 22, 23]. Poleg omenjenih korakov, ki služijo predvsem obdelavi in pripravi podatkov, pride za njimi še korak analize. Analiza je postopek, kjer raziskovalec uporabi različne metode in pristope nad zbranimi podatki z namenom tvorjenja uporabnih informacij ali odgovarjanja na vprašanja raziskave. V koraku analize se uporabljajo različne metode, kot so metode raziskovalne analitike (povprečja, histogrami, vizualizacije ...) [24] in bolj zahtevne metode (statistične metode, metode podatkovnega rudarjenja ...) med katere uvrščamo tudi metode napovedne analitike.

5.1 Zbiranje podatkov

Za glavni vir podatkov v tej raziskavi smo prejeli podatkovno bazo, ki vsebuje različne podatke o prodajah na spletnem prodajnem mestu podjetja Amazon. Podatke smo prejeli v obliki varnostne kopije relacijske podatkovne

baze MySQL od podjetja Sport2People, ki se ukvarja s prodajo na spletnem prodajnem mestu. Varnostno kopijo smo obnovili v prej za to pripravljen MySQL strežnik. Sledili sta analiza strukture podatkov in vsebinska analiza, s katerima smo identificirali podatke, ki nam prinašajo ustrezne informacije in tiste, ki jih ne potrebujemo.

Analiza strukture podatkov

Pri analizi strukture podatkov smo si pomagali s tvorjenjem diagrama ER (angl. entity–relationship) podatkovne baze. Za ta namen smo uporabili pripomoček "Reverse engineer" orodja MySQL Workbench, ki nam je ustvaril celoten diagram z vsemi 43 tabelami in povezavami med njimi [25]. Po tem smo z natančnim pregledom sheme začeli z izločanjem tabel, ki za razvoj naše rešitve ne nosijo koristnih informacij. To so večinoma sistemske tabele same aplikacije, ki uporablja podatkovno bazo. Izločili smo tudi tabele, ki jih aplikacija uporablja za koordinacijo elektronske pošte. Po izločanju smo dobili 17 tabel. V nadaljevanju smo z dodatnim filtriranjem poskušali ohraniti le strukturo podatkovne baze, ki predstavlja osnovne podatke, pomembne za napovedovanje kupcev in izdelkov. Odstranili smo tudi redundantne povezave, ki so bile tvorjene zgolj iz optimizacijskih razlogov. Diagram končne sheme je prikazan na sliki 5.1.

Sledi opis tabel podatkovne baze.

Marketplaces

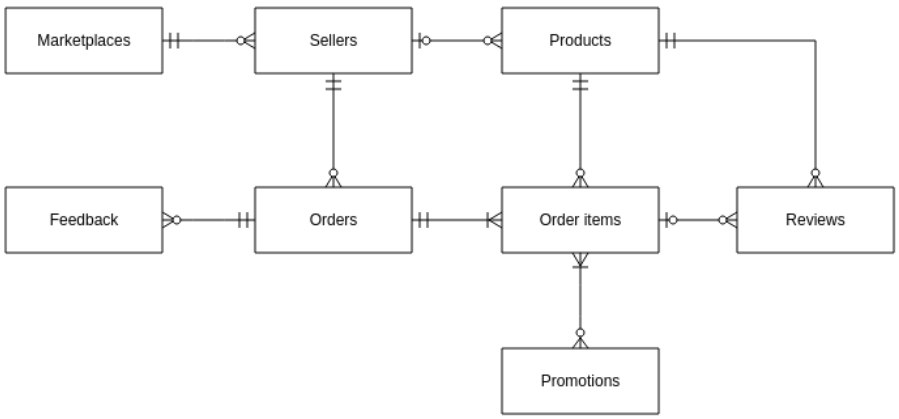
Tabela "Marketplaces" nosi podatke o posameznih Amazonovih prodajnih centrih (amazon.com, amazon.de, amazon.co.uk ...).

Sellers

Tabela "Sellers" hrani podatke o prodajalcih, ki prodajajo na spletnem prodajnem mestu Amazona. Pomembnejši stolpci podatkovne tabele "Sellers" so: ime prodajalne, e-poštni naslov, ključ za avtentikacijo idr.

Products

V tabeli "Products" se nahajajo podatki o izdelkih prodajalcev. Te po-



Slika 5.1: ER diagram z najbolj relevantnimi tabelami

datke je aplikacija pridobila s pomočjo Amazon MWS (Amazon Marketplace Web Service) API vmesnika. Stolpci tabele "Products" so naslednji: naslov, kratek naslov in ASIN (identifikacijska številka izdelka).

Feedback

"Feedback" je tabela odzivov kupcev za prodajalca. Vsak odziv je vezan na prodajalca oziroma njegovo spletno trgovino in vsebuje podatek o oceni (od 1 do 3) in besedilo odziva.

Orders

Tabela "Orders" vsebuje podatke o sklenjenih naročilih med različnimi kupci in prodajalci iz prej omenjene tabele "Sellers". Podatki o kupcih, ki so izvedli naročilo, niso ločeni, ampak so kar v tabeli naročil. Stolpci tabele "Orders", ki opisujejo kupca, so sledeči: ime kupca, e-poštni naslov kupca, telefonska številka in lokacijske informacije kupca. Ostali podatki o naročilu, hranjeni v tabeli "Orders", so: status naročila (neposlano, poslano, v pošiljanju, na čakanju, preklicano), podatki o dostavi (naslov, kontakt) in podatki o plačilu (plačilna valuta, naslov računa, datum plačila).

Order items

"Order items" je tabela, ki izpolnjuje relacijo mnogo proti mnogo med naročili in izdelki. V njej se hranijo ključi izdelkov in naročil ter še nekaj dodatnih informacij, kot so: količina naročenega izdelka, cena izdelka, ohranjenost izdelka (novo, rabljeno) in podatki o dajatvah (davek na dodano vrednost, cena dostave).

Reviews

V tabeli "Reviews" se hranijo podatki o pregledih, ki so jih kupci objavili za določen izdelek na spletnem prodajnem mestu Amazona. Tabela ima podatke o besedilu pregleda, oceni (od 1 do 5), naslovu pregleda in imenu kupca, ki je objavil pregled.

Promotions

Tabela "Promotions" hrani le en pomemben podatek, in sicer identifikacijsko številko promocije, ki jo je prodajalec aktiviral na spletnem prodajnem mestu Amazona. Promocije so povezane z izdelki iz naročil (tabela "Order items"), kar nam pove, kateri izdelki naročila so bili izvedeni v času katere od promocij.

Vsebinska analiza podatkov

Pri vsebinski analizi podatkov smo se osredotočili na vsebinski in statistični pregled podatkov. V tem delu analize smo pridobili informacije o količini in vplivnosti določenih podatkov, ki so nam kasneje prišle prav pri samem razvoju rešitve kot tudi pri pojasnitvah rezultatov.

Celotna varnostna kopija podatkovne baze vsebuje približno 12 GB podatkov. Od tega je največja tabela, ki hrani odposlano in elektronsko pošto v pošiljanju. Od tabel našega ožjega izbora iz prejšnjega razdelka je največja tabela "orders" s skoraj 3 GB podatkov. Za njo sledita tabela "order_items" z 2 GB podatkov in tretja največja "products", ki hrani samo 34 MB podatkov.

Osrednji del podatkov so naročila, ki povezujejo izdelke, kupce in prodajalce. Omenjene povezave nosijo ključno informacijo o zgodovini nakupov, ki jo bomo uporabljali pri napovedi kupcev in izdelkov. Podatki o naročilih so hranjeni v dveh podatkovnih tabelah: "orders" in "order_items". Prva tabela "orders" hrani informacije o naročilih in kupcih, medtem ko druga tabela "orders_items" tvori relacijo med naročilom in izdelki. Slednja tabela nam torej za vsako naročilo pove, kateri izdelki so bili naročeni. Cela podatkovna tabela "order_items" hrani približno 5 535 000 zapisov.

Vsak prodajalec na spletnem prodajnem mestu Amazona ima lahko več izdelkov, vendar ni nujno, da izdelek vedno pripada le enemu prodajalcu. Isti izdelek lahko prodaja več prodajalcev spletnega mesta Amazon. Kljub temu se je pri načrtovanju podatkovne baze izpustilo povezavo mnogo proti mnogo med izdelki in prodajalci. Namesto nje se v primeru izdelka, ki ima več prodajalcev, v tabelo "products" zapiše podvojen vnos. Poenostavitev je bila pri načrtovanju podatkovne baze smiselna zaradi majhne frekvence omenjenih primerov, enostavnosti in majhne verjetnosti, da bi se dva prodajalca, ki prodajata isti izdelek, pojavila v podatkovni bazi. Ponovljenih vnosov v tabeli izdelkov v podatkovni bazi je približno 1,7 %.

Kot tretji člen prodaje nastopi kupec. Kupci lahko na spletnem prodajnem mestu Amazona sklenejo več naročil oziroma nakupov. Med kupci in naročili torej velja povezava ena proti mnogo, vendar so razvijalci pri tvorjenju podatkovne baze ubrali podobno pot kot pri izdelkih in podatke o kupcih zasidrali kar v tabelo naročil. Najbolj pomemben podatek kupca je identifikacija, ki se v našem primeru hrani kot posredni elektronski poštni naslov kupca. Posredni elektronski poštni naslov je poseben, ker v samem naslovu ne vsebuje imena, vendar le niz naključnih črk. Omenjeni poštni naslov je nekakšen posrednik do pravega uporabnika. Pri pošiljanju elektronske pošte na posredni naslov kupec prejme sporočilo na svoj pravi elektronski poštni naslov. Za nas je ta poštni naslov pomemben, saj lahko preko njega povežemo naročila istega kupca. V podatkovni bazi najdemo nakupe s približno 4 240 000 različnimi kupci.

podatek/rezanje	Brez omejevanja	Samo poslovalnica US
prodajalci	497	386 (77,7%)
kupci	4 240 152	3 931 687 (92,7%)
izdelki	158 783	123 480 (77,8%)
naročila	5 535 119	5 110 763 (92,3%)
naročeni izdelki	5 834 395	5 393 750 (92,4%)

Tabela 5.1: Statistika vseh podatkov v primerjavi s podatki vezanimi na Amazon US poslovalnico

Na spletnem prodajnem mestu podjetja Amazon lahko prodajalec pospešuje prodajo s tvorjenjem različnih promocij. V podatkovni bazi se v primeru, da je bil nek izdelek kupljen v promociji, hrani identifikacijska številka promocije. Statistična analiza podatkovne baze nam govori, da je bilo 33,2 % naročil izdelkov naročenih v okviru promocije.

Če je kupcu naročeni izdelek všeč ali ne, lahko svoje izkušnje z izdelkom pusti pod tako imenovanimi pregledi (angl. reviews). V podatkovni bazi se nahaja 13 201 zapisov pregledov, ki so vezani na 101 izdelkov, kar je samo 0,06 % vseh izdelkov v podatkovni bazi. Zaradi premajhne količine podatkov o pregledih ne bomo uporabili v razvoju rešitve.

Podjetje Amazon ima po svetu več poslovalnic. Če pogledamo bolj natančno, jih je imel v času pisanja magistrskega dela 11. Od 11 poslovalnic podatkovna baza hrani podatke iz devetih različnih poslovalnic, s poudarkom na največji Amazonovi poslovalnici na ameriškem trgu. Ker je večina podatkov z ameriškega trga domene ".com", smo tudi mi za raziskavo uporabili zgolj podatke iz ameriške poslovalnice Amazon US. S to omejitvijo smo glavno množico naročil zmanjšali za samo 7,7 %. Dodatne statistične primerjave si lahko pogledate v tabeli 5.1.

Zbiranje podatkov iz dodatnega vira

Podatki iz glavnega vira (podatkovne baze) kljub veliki količini nimajo velike širine. S širino podatkov mislimo na več dodatnih informacij posameznih podatkovnih entitet, predvsem izdelkov.

Odločili smo se, da bomo poskušali pridobiti dodatne podatke o izdelkih s spletnega prodajnega mesta Amazona. Za dostop do Amazonovih storitev pri prodaji obstaja več aplikacijskih programskih vmesnikov (API). Poglejmo si opis dveh najbolj uporabljenih aplikacijskih programskih vmesnikov Amazona:

Amazon Marketplace Web Service (MWS)

API MWS je namenjen prodajalcem spletnega prodajnega mesta in omogoča programski dostop do podatkov o izdelkih, njihovih prodajnih ponudbah, naročilih ter omogoča tvorjenje različnih poročil. Avtentikacija se izvaja na nivoju posameznega prodajalca preko dveh žetonov. Prvi žeton je žeton aplikacije, imenovan tudi žeton dostopa (angl. access token). Vsaka aplikacija ima svoj žeton dostopa in ko prodajalec želi določeni aplikaciji omogočiti dostop, mora na spletnem mestu Amazona tvoriti drug avtorizacijski žeton (angl. authorization token), ki ga potem posreduje aplikaciji. Aplikacija lahko s kombinacijo omenjenih žetonov pridobiva in uporablja podatke, do katerih lahko drugače dostopa le sam prodajalec.

Amazon Product Advertising (APA)

Aplikacijski vmesnik APA je namenjen tvorjenju oglasov in pridobivanju različnih javnih podatkov izdelkov. Tukaj se avtentikacija izvaja le nad aplikacijo, ki uporablja API. Za izvajanje akcij aplikacija potrebuje 3 ključe. Podatki, ki jih lahko pridobimo iz API-ja APA, so po večini podatki, ki jih lahko najdemo na sami spletni strani spletnega prodajnega mesta Amazona.

Pri naši raziskavi nismo potrebovali natančnih podatkov in poročil od specifičnega prodajalca, vendar le nekaj dodatnih podatkov o vseh izdelkih iz

glavnega vira oziroma podatkovne baze. Za ta namen je dovolj, da uporabimo samo API APA. S pomočjo poizvedbe za izdelke vmesnika APA smo lahko pridobili dodatne informacije o izdelkih, kot so trenutna cena, kategorija in rang izdelka. Rang izdelka je ocena pozicije izdelka v primerjavi z ostalimi izdelki iz kategorij, v katere je izdelek opredeljen.

5.2 Procesiranje podatkov

Preden smo podatke uporabili pri testih, smo jih preoblikovali in prečistili. Do podatkov v podatkovni bazi MySQL smo dostopali s pomočjo pythonske knjižnice pmysql. Vse podatke smo skozi analizo predstavili kot tabelo naročenih izdelkov, kjer vsaka vrstica tabele predstavlja en naročen izdelek iz nekega naročila. Vsak izdelek v naročilu je bil lahko naročen v večjih količinah, kar je zapisano v stolpcu količina (angl. quantity). Ostali stolpci, ki smo jih pridobili iz podatkovne baze MySQL so opisani v nadaljevanju:

- prodajalec ... identifikacija prodajalca,
- kupec ... identifikacija kupca,
- izdelek ... identifikacija izdelka,
- ocena ... povprečna ocena pregledov za izdelek,
- promocija ... identifikacija promocije, v kateri je bil izdelek naročen (če ni bil v promociji, je vrednost "None"),
- količina ... količina naročenih izdelkov.

Čiščenje podatkov

Naročila kupcev, ki so naročali manj kot trinajstkrat, smo izločili. S tem smo se znebili naročil in kupcev, za katere je skoraj nemogoče napovedati njihova prihodnja naročila iz preteklih naročil, saj jih imajo premalo. Prav tako, smo podatke dovolj zmanjšali za sorazmerna hitra testiranja. Omejitev

trinajst je bila izbrana na podlagi minimalne možne omejitve, pri kateri je bilo mogoče pognati teste v danem testnem okolju.

Opazili smo, da nekatera naročila, pridobljena iz podatkovne baze, nimajo podatkov o kupcih. Ker za tovrstna naročila ne moremo vedeti, kateri kupec jih je naročil, smo jih izločili iz tabele naročenih izdelkov.

Združevanje virov

Dodatne podatke o izdelkih, pridobljene iz vmesnika Amazon APA, smo združili s pomočjo identifikacijskega ključa izdelkov (ASIN). Uporabili smo pythonsko knjižnico za API APA, ki ima že implementirane in poenostavljene API-poizvedbe. Ker smo si želeli dodatnih podatkov o izdelkih, smo v dokumentaciji API-vmesnika poiskali primerno poizvedbo [26]. Ta poizvedba je "ItemLookup". Ob tvorjenju poizvedbe "ItemLookup" podamo identifikacijski ključ izdelka in kot rezultat prejmemo podrobnejše informacije o izdelku. Za hitrejše poizvedovanje nam API omogoča množične akcije, s katerimi lahko izvedemo do 5 enakih poizvedb hkrati. Tako smo ASIN-ključe izdelkov najprej združili po 5, nato pa s pomočjo knjižnice pmysql izvedli poizvedbe. S pomočjo vmesnika APA nam je tabelo podatkov uspelo razširiti s sledečimi stolpci:

- kategorija ... oznaka kategorije, v kateri se nahaja izdelek,
- cena ... trenutna cena izdelka,
- rang ... rang izdelka.

Iz dodatnih stolpcev lahko poizvemo po dodatnih zanimivih statističnih podatkih. Povprečno število kategorij, v katerih posamezen kupec kupuje, je enako 1,7. Povprečna cena naročenih izdelkov je enaka 28 evrov. Povprečen rang izdelkov, ki jih kupujejo kupci, pa je 34 245, kar je presenetljivo visoko. Vzrok za tako visok rang je, da so naročila iz podatkovne baze več mesecev starejša od podatkov, pridobljenih s pomočjo API-vmesnika APA. Od takrat so se rangi večine izdelkov lahko zelo spremenili. Podatkov o ceni in rangu izdelkov se bomo zato v razvoju pristopov izogibali.

"Execution is so critical. Sometimes you just need to try something, see what works and move forward."

— Meenal Balar

Poglavje 6

Eksperimentalno ovrednotenje

V zadnjem vsebinskem poglavju bomo predstavili testno okolje in rezultate eksperimentalnega ovrednotenja pristopov, ki smo jih razvili v sklopu magistrskega dela. Rezultate naših pristopov bomo primerjali z rezultati primerjalnih metod, ki jih bomo definirali v nadaljevanju.

6.1 Primerjalne metode

Uspešnost naše rešitve bomo ugotavljali s pomočjo primerjav s petimi primerjalnimi metodami. Vse primerjalne metode, tako kot tudi naša rešitev, implementirajo model napovedovanja, ki smo ga definirali v poglavju 3. Potemtakem velja, da bodo rešitve vseh metod ocenjene z istimi ocenjevalnimi kriteriji in zaradi tega primerljive med sabo. Primerjalne metode, definirane v nadaljevanju, bodo predstavljene v domeni napovedovanja kupcev za izdelek $\boldsymbol{i}$, vendar veljajo tudi za napovedovanje izdelkov za poljubnega kupca $\boldsymbol{b}$.

Naključna metoda

Naključna metoda nam bo pri primerjavi rezultatov služila kot spodnja meja sprejemljivih napovedi. Zanjo velja, da je enostavna in hitra.

Poglejmo si postopek, kako naključna metoda napoveduje kupce za izdelek $\boldsymbol{i}$. Dodaten vhodni parameter naključne metode je velikost napovednega vzorca $\boldsymbol{k}$. Število $\boldsymbol{k} \in [0, 100]$ določa delež velikosti napovedne množice $\boldsymbol{A_i}$ glede na velikost celotne množice kupcev $\boldsymbol{\mathcal{B}}$. Napovedno množico $\boldsymbol{A_i}$ sestavimo z izbiranjem naključnih kupcev iz množice $\boldsymbol{\mathcal{B}}$, pri čemer ne izbiramo kupcev, ki so že v napovedni množici $\boldsymbol{A_i}$. Popolnoma analogen postopek je pri napovedih izdelkov za kupca $\boldsymbol{b}$, le nad drugimi množicami.

Metoda kategorij

Metoda kategorij je izboljšava naključne metode. Pri tej metodi za napovedovanje uporabimo informacijo o kategorijah izdelkov. Implementacija metode kategorij je za napovedovanje kupcev drugačna kot za napovedovanje izdelkov, zato ju bomo predstavili ločeno.

Algoritem 15 Metoda kategorij

```
F ← ∅
c ← KATEGORIJA(i)
for all (b, I_o) ∈ O do
    for all i_1 ∈ I_o do
        if c = KATEGORIJA(i_1) then
            if (b : freq) ∈ F then
                freq ← freq + 1
            else
                F ← F ∪ (b : 1)
A ← ∅
for all (b : freq) ∈ F do
    if freq ≥ k then
        A ← A ∪ b
```

Metoda kategorij za napovedovanje kupcev Ideja pri napovedovanju kupcev z metodo kategorij je, da so napovedani kupci $\boldsymbol{A_i}$ izdelka $\boldsymbol{i}$ tisti

kupci, ki največ naročajo v kategoriji izdelka $\boldsymbol{i}$. Tudi tukaj imamo vhodni parameter $\boldsymbol{k}$, ki pa ima drugačen pomen, kot pri naključni metodi. Število $\boldsymbol{k}$ določa spodnjo mejo frekvence naročil kupcev, ki jih napovedujemo za izdelek $\boldsymbol{i}$. To pomeni, da bodo kupci, ki so v kategoriji izdelka $\boldsymbol{i}$ izvedli več od ali enako kot $\boldsymbol{k}$ naročil, dodani v napovedno množico $\boldsymbol{A_i}$.

Metoda je sestavljena iz dveh delov. V prvem delu za kupce, ki so naročali izdelke iz kategorije izdelka $\boldsymbol{i}$, določimo frekvenco naročil, ki jih uporabimo v drugem delu za tvorjenje napovedne množice $\boldsymbol{A_i}$. Podroben postopek je zapisan s psevdokodo v algoritmu 15.

Metoda kategorij za napovedovanje izdelkov Pri napovedovanju izdelkov je metoda kategorij zelo podobna naključni metodi, saj temelji na enakem postopku naključne izbire iz množice izdelkov. Tokrat je množica izdelkov, iz katere izbiramo, omejena, in sicer s kategorijo, iz katere je kupec $\boldsymbol{b}$ naročal največ.

Postopek napovedovanja izdelkov za kupca $\boldsymbol{b}$ z metodo kategorij je sledeč. V prvem delu iteriramo preko vseh naročil kupca $\boldsymbol{b}$ in preštejemo, kolikokrat je naročal v posamezni kategoriji. V drugem delu tvorimo množico izdelkov iz kategorije, v kateri je kupec $\boldsymbol{b}$ naročal največ izdelkov. Iz dobljene množice naključno izberemo $\boldsymbol{k}$ izdelkov, ki tvorijo napovedno množico $\boldsymbol{C}$.

Regresijske metode

Regresijske metode so najbolj znane in najpogosteje uporabljene metode za napovedovanje. Prav zato smo nekaj bolj znanih regresijskih metod uporabili za primerjavo z rešitvijo, ki smo jo razvili. Regresijske metode uvrščamo med metode statističnega modeliranja, ki so zelo pogosto uporabljene pri podatkovnem rudarjenju. Več o regresijskih metodah se nahaja v [27, 28], tukaj pa se omejimo izključno na njihovo uporabo.

Namen regresijskih metod je, da ocenijo relacijo med odvisno in neodvisnimi spremenljivkami. Iz dobljene relacije lahko iz neodvisne spremenljivke kasneje ocenimo odvisno spremenljivko. Prvi del, kjer ocenjujemo relacijo

med spremenljivkami, imenujemo učenje. Drugi del, kjer ocenjujemo odvisno spremenljivko, imenujemo napovedovanje.

Neodvisne spremenljivke so običajno predstavljene z matriko $\boldsymbol{X}$, kjer so vrstice primeri in stolpci spremenljivke primerov. Odvisna spremenljivka je predstavljena z vektorjem $\boldsymbol{y}$, kjer vsaka vrednost v vektorju predstavlja oceno oziroma napoved primera.

Pri napovedovanju kupcev za izdelek $\boldsymbol{i}_a$ se v vrsticah matrike $\boldsymbol{X}$ za posameznega $\boldsymbol{b} \in \boldsymbol{\mathcal{B}}$ nahajajo števila naročenih izdelkov. Vrednost v celici $\boldsymbol{x}_{\boldsymbol{u},\boldsymbol{v}}$ predstavlja, koliko izdelkov $\boldsymbol{i}_{\boldsymbol{v}}$ je naročil kupec $\boldsymbol{b}_{\boldsymbol{u}}$. Vektor $\boldsymbol{y}$ predstavlja napovedane količine izdelka $\boldsymbol{i}_a$ za posameznega kupca $\boldsymbol{b}$.

Odvisna in neodvisna spremenljivka pri napovedovanju kupcev za izdelek $\boldsymbol{i}_a$ sta:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_{1,1} & \boldsymbol{x}_{1,2} \dots & \boldsymbol{x}_{1,k} \\ \boldsymbol{x}_{2,1} & \boldsymbol{x}_{2,2} \dots & \boldsymbol{x}_{2,k} \\ \vdots & & \\ \boldsymbol{x}_{n,1} & \boldsymbol{x}_{n,2} \dots & \boldsymbol{x}_{n,k} \end{bmatrix}, \boldsymbol{y} = \begin{bmatrix} \boldsymbol{x}_{1,a} \\ \boldsymbol{x}_{2,a} \\ \vdots \\ \boldsymbol{x}_{n,a} \end{bmatrix},$$

pri čemer je $n = |\boldsymbol{\mathcal{B}}|$, $k = |\boldsymbol{\mathcal{I}}| - 1$ in $\boldsymbol{x}_{\boldsymbol{u},\boldsymbol{v}}$ količina naročenih izdelkov $\boldsymbol{i}_{\boldsymbol{v}}$, ki jih je kupec $\boldsymbol{b}_{\boldsymbol{u}}$ naročil v preteklosti.

Za primerjavo smo uporabili naslednje regresijske metode:

- linearna regresija,
- logistična regresija,
- regresija k-najbližjih sosedov (KNN-regresija),
- regresija naključnih gozdov (RF-regresija).

Linearna regresija je ena od najenostavnejših in najpogosteje uporabljenih regresijskih metod. Pri njej je model napovedovanja premica, ki jo želimo v postopku učenja definirati tako, da se čim bolj prilega vsem točkam oziroma primerom modela. Pri napovedovanju iz neodvisnih spremenljivk in ocenjenih parametrov premice izračunamo odvisno spremenljivko.

Logistična regresija je regresijska metoda, kjer je nabor vrednosti odvisne spremenljivke logičen $\{0, 1\}$. Model, ki ga optimiziramo, temelji na logistični funkciji.

Regresija $\boldsymbol{k}_{knn}$ najbližjih sosedov (KNN-regresija) temelji na razdaljah med primeri $\boldsymbol{x}$. Pri napovedovanju regresijska metoda KNN poišče $\boldsymbol{k}_{knn} \in \mathbb{N}$ najbližjih sosedov iz učne množice in povpreči njihove pripadajoče odvisne spremenljivke $\boldsymbol{y}$.

Regresija naključnih gozdov je ena izmed ansambelskih oziroma sestavljenih metod, ki temelji na odločilnih drevesih. Naključni gozdovi vpeljujejo izboljšavo, ki odpravlja pojav prekomernega prilagajanja testnim podatkom [29].

Pri vseh regresijskih metodah, razen pri logistični regresiji, je rezultat vektor $\boldsymbol{y} \in \mathbb{R}^{|\boldsymbol{\mathcal{B}}|}$, ki za vsakega kupca predstavlja napovedano količino izdelka $\boldsymbol{i}$. Ker so vrednosti v vektorju pozitivna realna števila, moramo uvesti mejo, ki bo določala, kateri kupci so potencialni kupci in kateri ne. Omenjena meja je vhodni parameter $\boldsymbol{k} \in [0, 1]$. Če je napovedana količina izdelka $\boldsymbol{i}$ za kupca $\boldsymbol{b}$ v vektorju $\boldsymbol{y}$ večja od $\boldsymbol{k}$, potem za kupca $\boldsymbol{b}$ napoveduje, da bo ta naročil izdelek $\boldsymbol{i}$ in velja $\boldsymbol{i} \in \boldsymbol{A_i}$. Če je napovedana količina manjša od $\boldsymbol{k}$, potem metoda za kupca $\boldsymbol{b}$ napoveduje, da ne bo naročil izdelka $\boldsymbol{i}$.

6.2 Testno okolje

Teste smo poganjali na računalniškem sistemu s štirijedrnim procesorjem Intel Core i7-6700HQ, kjer lahko takt vsakega jedra doseže do 3,2 GHz. Predpomnilnik velikosti 6 MB skrbi za hitrejše delo procesorja z delovnim pomnilnikom RAM, velikim 20 GB. Podatke hranimo na disku PCIe SSD, velikem 256 GB. Računalniški sistem uporablja 64-bitni operacijski sistem Arch Linux z jedrom različice 4.12.10.

Za implementacijo pristopov in testov smo uporabili programsko okolje Python 3 [30] z naslednjimi programskimi knjižnicami:

- networkx ... knjižnica za delo z grafi,

- pymysql ... knjižnica za povezavo in delo s podatkovno bazo MySQL,
- sklearn ... python zbirka orodij za podatkovno rudarjenje,
- scipy ... knjižnica za napredno delo z numeričnimi podatki.

Med razvojem smo za upravljanje z izvorno kodo uporabljali sistem Git. Repozitorij implementacije pristopov in testov je objavljen v [31].

Podatki, uporabljeni v testih so naročila iz podatkov, ki smo jih predstavili v poglavju 5. Naročila smo razdelili na učno (80 % celotne množice) in testno množico (20 % celotne množice) naročil. Učna množica je v testih uporabljena kot vhodna množica za učenje (gradnja grafov oziroma učenje regresijskih metod). Testna množica pa je bila uporabljena za napovedovanje in ocenjevanje natančnosti napovedi. Ocene vseh testov smo tvorili glede na ocenjevalne kriterije, ki smo jih definirali v poglavju 3.

Za ovrednotenje oziroma ocenjevanje metod se pri podatkovnem rudarjenju velikokrat uporablja tako imenovana metoda prečnega preverjanja [32, Poglavje 5.3], vendar je mi nismo uporabili, ker so naši podatki časovno odvisni. Naročila kupcev v prihodnosti sledijo naročilom, ki jih je kupec izvajal v preteklosti in ne obratno. Zato je pomembno, da testno množico sestavljajo naročila, ki jih je kupec izvedel kasneje kot naročila iz učne množice.

6.3 Ovrednotenje napovedovanja kupcev

Najprej bomo predstavili rezultate primerjalnih metod pri napovedovanju kupcev, na koncu pa jih bomo primerjali z rezultati v magistrskem delu razvitih pristopov.

Preden nadaljujemo, poglejmo povzetek ocenjevalnih kriterijev pri napovedovanju kupcev:

- 1. ocenjevalni kriterij $\overline{T^+/B}$ je povprečen delež pravilno napovedanih kupcev proti vsem napovedanimi kupcem,

- 2. ocenjevalni kriterij $\overline{T^+/A}$ je povprečen delež pravilno napovedanih kupcev proti dejanskim kupcem,

- 3. ocenjevalni kriterij $\overline{T^+/(A \cup B)}$ je povprečen delež pravilno napovedanih kupcev proti kupcem, ki so napovedani ali dejanski,

- 4. ocenjevalni kriterij $\overline{(T^+ - \alpha F^+)/A}$ je povprečen dobiček na napovedanega kupca, katerega razmerje med stroški in prihodki določa α.

V tabelah ocen bo poleg ocenjevalnih kriterijev informativno prikazan dodaten stolpec $\overline{A/\mathcal{B}}$, ki prikazuje delež napovedne množice glede na vso populacijo kupcev.

Ocena dobička oziroma četrtega ocenjevalnega kriterija je odvisna od α in zato brez določitve le-te pri testih ne moremo oceniti dobička. Oceno α bomo določili z naslednjim primerom.

Recimo, da napovedi kupcev za izdelek i uporabljamo za pošiljanje promocijske ponudbe za izdelek i po elektronski pošti napovedanim kupcem. Pošiljanje ponudbe kupcu, ki ga izdelek i ne zanima, ne bo imel vedno negativnega vpliva. Vendar recimo, da se bo vsak 100. kupec, ki ga izdelek i ne zanima, odjavil z našega elektronskega poštnega seznama. Seveda smo s tem, ko smo izgubili kupca, izgubili tudi recimo povprečno 5 naročil, ki bi jih izgubljeni kupec v prihodnosti opravil. Ocena α določa, kako močno napačno napovedan kupec vpliva na dobiček proti pravilno napovedanemu kupcu, in tako lahko iz primera ocenimo $\alpha = \frac{1}{100} * 5 = 0{,}05$. Pri tem prvi faktor določa pogostost dogodka (vsak 100. kupec), drugi pa moč negativnega dogodka (10 izgubljenih nakupov). Vse ocene dobička, prikazane v nadaljevanju, so za lažjo berljivost in večjo natančnost pomnožene s faktorjem 100.

Naključna metoda

Povzetek ocen rezultatov naključne metode je predstavljen v tabeli 6.1. Vsaka vrstica tabele predstavlja ocene za vhodni parameter k. Ocene rezultatov pri vseh vhodnih parametrih k so prikazane v dodatku A.

k	$\overline{T^+/B}$	$\overline{T^+/A}$	$\overline{T^+/(A \cup B)}$	Dobiček	$\overline{A/\mathcal{B}}$
100	100,00 %	0,76 %	0,76 %	−4,2	100,00 %
93	91,35 %	0,76 %	0,76 %	−4,2	93,00 %
6	4,64 %	0,74 %	0,50 %	−4,22	5,99 %
0	0,00 %	0,00 %	0,00 %	0	0,00 %

Tabela 6.1: Ocene napovedovanja kupcev naključne metode

Pri $k = 100$, kjer je napovedna množica največja, je $A = \mathcal{B}$, kar pomeni, da je metoda napovedala vse kupce. Pri tem je naključna metoda pričakovano odkrila vse dejanske kupce iz množice B, vendar pa so vsi ostali kupci napovedani napačno, kar pomeni, da je velikost množice F^+ maksimalna. Dobiček pri $\alpha = 0{,}05$ sega v negativno vrednost, kar pomeni, da bi lahko v primeru promocije, kjer bi uporabili naključno metodo, na dolgi rok pridelali negativen vpliv.

Pri vrednosti vhodnega parametra $k = 0$, kjer je napovedna množica najmanjša oziroma enaka 0, nismo napovedali nobenega kupca. V tem primeru se ni zgodilo nič in prav tako je dobiček enak 0.

Pri vmesnih vrednostih vhodnega parametra k opazimo, da naključna metoda v bistvu ne izboljša rezultatov, saj ocena tretjega ocenjevalnega kriterija $\overline{T^+/(A \cup B)}$ z manjšanjem k pada in dobiček ostaja negativen.

Metoda kategorij

Metoda kategorij je po tretjem ocenjevalnem kriteriju rezultate naključne metode izboljšala na 6,4 % pri $k = 7$. Pri tem je odkrila 82,8 % dejanskih kupcev. Da je metoda boljša od naključne, je razvidno tudi iz stolpca dobička v tabeli 6.2, ki je v najboljšem primeru zrasel do vrednosti 2,3.

Regresijske metode

Metode regresije so občutno izboljšale rezultate v primerjavi z naključno metodo in metodo kategorij.

k	$\overline{T^+/B}$	$\overline{T^+/A}$	$\overline{T^+/(A \cup B)}$	Dobiček	$\overline{A/\mathcal{B}}$
0	94,79 %	4,81 %	4,76 %	0,06	42,64 %
6	84,57 %	6,75 %	6,30 %	2,16	34,07 %
7	82,82 %	6,88 %	6,43 %	2,3	33,41 %
8	80,17 %	6,42 %	5,89 %	1,86	32,41 %
14	31,33 %	5,27 %	4,32 %	0,74	9,72 %
15	26,07 %	5,14 %	4,28 %	0,6	7,88 %

Tabela 6.2: Ocene napovedovanja kupcev z metodo kategorij

k	$\overline{T^+/B}$	$\overline{T^+/A}$	$\overline{T^+/(A \cup B)}$	Dobiček	$\overline{A/\mathcal{B}}$
0	85,16 %	1,01 %	0,99 %	−3,94	80,60 %
0,07	45,95 %	10,42 %	9,52 %	6,38	14,69 %
0,13	35,56 %	10,54 %	9,36 %	6,61	8,57 %
0,93	8,36 %	5,11 %	3,94 %	1,36	1,28 %
1	7,87 %	4,69 %	3,65 %	0,96	1,24 %

Tabela 6.3: Ocene napovedovanja kupcev z linearno regresijo

Ocene rezultatov linearne regresije v tabeli 6.3 nakazujejo, da se je linearna regresija najbolje odrezala pri vrednostih vhodnega parametra k blizu 0. Pri vhodnem parametru $k = 0{,}07$ imamo najvišjo oceno tretjega ocenjevalnega kriterija, medtem ko je dobiček največji pri $k = 0{,}13$ zaradi manjše napovedne množice in posledično manjše množice F^+.

k	$\overline{T^+/B}$	$\overline{T^+/A}$	$\overline{T^+/(A \cup B)}$	Dobiček	$\overline{A/\mathcal{B}}$
0	100,00 %	0,76 %	0,76 %	−4,2	100,00 %
0,07	54,38 %	9,84 %	8,08 %	6,16	15,85 %
0,13	49,63 %	10,40 %	7,84 %	6,86	15,26 %
0,93	36,45 %	7,41 %	3,91 %	4,75	14,23 %
1	36,45 %	7,42 %	3,92 %	4,75	14,21 %

Tabela 6.4: Ocene napovedovanja kupcev z RF-regresijo

Podobno kot linearna regresija ima RF-regresija najboljši rezultat pri nizkem vhodnem parametru $\boldsymbol{k}$. Kot je iz stolpca $\overline{\boldsymbol{A}/\boldsymbol{\mathcal{B}}}$ tabele 6.4 razvidno, je RF-regresija pri najboljšem rezultatu vrnila večjo napovedno množico kot linearna regresija. Kljub temu je linearna regresija glede na tretji ocenjevalni kriterij malo boljša od RF-regresije.

$\boldsymbol{k}$	$\overline{\boldsymbol{T}^+/\boldsymbol{B}}$	$\overline{\boldsymbol{T}^+/\boldsymbol{A}}$	$\overline{\boldsymbol{T}^+/(\boldsymbol{A}\cup\boldsymbol{B})}$	Dobiček	$\overline{\boldsymbol{A}/\boldsymbol{\mathcal{B}}}$
0	100,00 %	0,76 %	0,76 %	−4,2	100,00 %
0,27	68,85 %	3,86 %	3,35 %	0,1	47,82 %
0,33	65,88 %	4,04 %	3,46 %	0,36	46,80 %
0,47	62,41 %	4,01 %	3,19 %	0,42	45,83 %
0,93	48,92 %	2,80 %	2,25 %	−0,28	41,15 %
1	48,92 %	2,80 %	2,25 %	−0,28	41,15 %

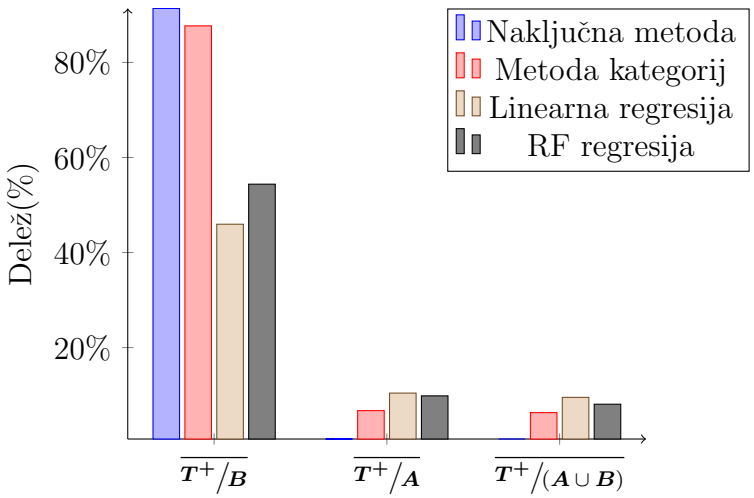
Tabela 6.5: Ocene napovedovanja kupcev s KNN-regresijo

KNN-regresija se je za napovedovanje kupcev izkazala za zelo slabo metodo, saj je slabša od metode kategorij; boljša je le od naključne metode. Ocene napovedi KNN-regresije pri $\boldsymbol{k}_{knn} = 3$ so prikazane v tabeli 6.5. Ocene rezultatov pri drugih vrednostih $\boldsymbol{k}_{knn}$ so podobne ali slabše, zato so izpuščene.

$\overline{\boldsymbol{T}^+/\boldsymbol{B}}$	$\overline{\boldsymbol{T}^+/\boldsymbol{A}}$	$\overline{\boldsymbol{T}^+/(\boldsymbol{A}\cup\boldsymbol{B})}$	Dobiček	$\overline{\boldsymbol{A}/\boldsymbol{\mathcal{B}}}$
2,74 %	10,73 %	1,82 %	9,47	0,18 %

Tabela 6.6: Ocene napovedovanja kupcev z logistično regresijo

Ocene rezultatov logistične regresije so prikazane v tabeli 6.6. Logistična regresija vedno vrača logično vrednost, zato tukaj nimamo vhodnega parametra $\boldsymbol{k}$, ki bi določal mejo odločanja. Logistična regresija je, po tretjem ocenjevalnem kriteriju sodeč, slabša od KNN-regresije, vendar ima boljši dobiček na napovedanega kupca. V bistvu ima dobiček boljši tudi od linearne regresije in RF-regresije, vendar logistična regresija tvori veliko manjšo napovedno množico od ostalih regresijskih metod, kar pomeni, da napove zelo majhno število kupcev.



Slika 6.1: Diagram prvih treh ocenjevalnih kriterijev

Na sliki 6.1 je stolpčni diagram, ki primerja najboljše ocene prvih treh kriterijev naključne metode, metode kategorij, linearne regresije in RF-regresije. Iz diagrama je razvidno, da je linearna regresija najboljša po drugem $\overline{T^+/A}$ in tretjem ocenjevalnem kriteriju $\overline{T^+/(A \cup B)}$. Takoj za njo je RF-regresija, ki pa v primerjavi z linearno regresijo odkrije več dejanskih kupcev $\overline{T^+/B}$. Največ dejanskih kupcev odkrije naključna metoda, ker za napovedno množico vzame velik delež vseh kupcev, vendar je iz drugega ocenjevalnega kriterija $\overline{T^+/B}$ razvidno, da ima naključna metoda največji delež napačno napovedanih kupcev.

Pristop 1: Skupni sosedje

Teste za napovedovanje kupcev pri obeh naših pristopih smo izvajali za vse kombinacije vhodnih parametrov funkcij uteži:

- utež 1 ... število skupnih izdelkov,
- utež 2 ... število skupnih izdelkov z naročenimi količinami,
- utež 3 ... skupne kategorije

in omejevanja grafa: $\boldsymbol{k} \in [0, 1, 2, \ldots 15]$.

Ocene rezultatov prvega pristopa naše rešitve kažejo na podobno uspešnost kot uspešnost metode kategorij. Metoda prvega pristopa je bila najboljša pri uteži 1 in vhodnem parametru $\boldsymbol{k} = 1$. Izkazalo se je, da pri vseh treh utežeh prevladujejo rezultati z vhodnim parametrom $\boldsymbol{k} = 1$.

Pri vhodnem parametru $\boldsymbol{k} = 0$ je za napovedovanje uporabljen poln graf kupcev, saj ga ne omejujemo. Vendar delež napovedne množice $\overline{\boldsymbol{A}/\boldsymbol{\mathcal{B}}}$ ne dosega 100 %. Tukaj se pojavi vprašanje, zakaj je tako, če so po predpostavki polnega grafa sosedje kateregakoli kupca v grafu vsi ostali sosedje. Slednje je posledica dveh vzrokov. Prvi je razlika med testno in učno množico. Kupci, ki so v učni množici, ne nastopajo nujno tudi v testni množici in obratno. Zato se pojavi razlika v maksimalnem številu vseh kupcev v napovedni množici in vseh kupcev vseh naročil. Drugi vzrok se skriva pri posebnosti prvega pristopa, pristopa skupnih sosedov. V tem pristopu za združevanje vseh sosedov kupcev uporabljamo presek, kar je razvidno iz algoritma 11. Ker v preseku zajamemo le sosede $\boldsymbol{\mathcal{N}}(\boldsymbol{b})$ brez kupca $\boldsymbol{b}$, bo slednji zagotovo ostal izven napovedne množice. Ker je drugi vzrok, ki smo ga ravnokar opisali, značilen le za prvi pristop, bomo pri drugem pristopu za $\boldsymbol{k} = 0$ opazili večji delež napovedanih kupcev med vsemi kupci $\overline{\boldsymbol{A}/\boldsymbol{\mathcal{B}}}$.

$\boldsymbol{k}$	$\overline{\boldsymbol{T^+}/\boldsymbol{B}}$	$\overline{\boldsymbol{T^+}/\boldsymbol{A}}$	$\overline{\boldsymbol{T^+}/(\boldsymbol{A} \cup \boldsymbol{B})}$	Dobiček	$\overline{\boldsymbol{A}/\boldsymbol{\mathcal{B}}}$
0	37,58 %	0,48 %	0,47 %	−4,2	92,09 %
1	12,38 %	7,64 %	4,61 %	5,83	0,68 %
2	6,16 %	4,91 %	2,80 %	3,81	0,23 %
14	0,06 %	0,49 %	0,06 %	0,38	0,01 %
15	0,00 %	0,08 %	0,00 %	−0,05	0,01 %

Tabela 6.7: Ocene napovedovanja kupcev s prvim pristopom in utežjo 1

Iz povzetka tabele 6.7 je razvidno, da najboljša ocena prve uteži po tretjem ocenjevalnem kriteriju tvori slabši dobiček (4,6 %) kot linearna regresija, vendar boljšega kot logistična in KNN-regresija.

Ocene druge uteži v tabeli 6.8 pri $\boldsymbol{k} = 1$ kažejo na enako uspešnost.

k	$\overline{T^+/B}$	$\overline{T^+/A}$	$\overline{T^+/(A \cup B)}$	Dobiček	$\overline{A/\mathcal{B}}$
0	37,58 %	0,48 %	0,47 %	−4,2	92,09 %
1	12,38 %	7,64 %	4,61 %	5,83	0,68 %
2	8,40 %	6,02 %	3,41 %	4,44	0,35 %
14	0,68 %	1,44 %	0,51 %	1,13	0,02 %
15	0,49 %	0,90 %	0,34 %	0,56	0,02 %

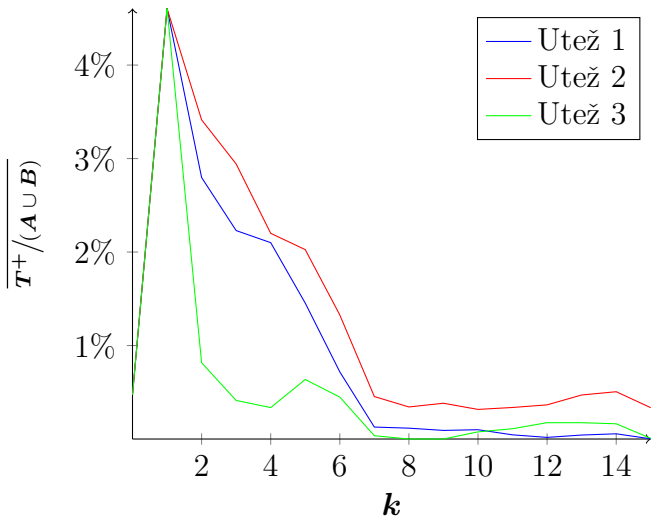
Tabela 6.8: Ocene napovedovanja kupcev s prvim pristopom in utežjo 2

k	$\overline{T^+/B}$	$\overline{T^+/A}$	$\overline{T^+/(A \cup B)}$	Dobiček	$\overline{A/\mathcal{B}}$
0	37,58 %	0,48 %	0,47 %	−4,2	92,09 %
1	12,38 %	7,64 %	4,61 %	5,83	0,68 %
2	2,23 %	3,26 %	0,82 %	2,51	0,15 %
14	0,16 %	0,99 %	0,16 %	0,85	0,01 %
15	0,01 %	0,25 %	0,01 %	0,22	0,01 %

Tabela 6.9: Ocene napovedovanja kupcev s prvim pristopom in utežjo 3

Podobno pri uteži 3 v tabeli 6.9. Izkaže se, da je najvplivnejši faktor za napovedovanje kupcev že sama povezava med kupcema, ki narekuje, ali imata dva kupca enega ali več skupnih izdelkov. Zaradi tega imajo vse tri ocene enak in najboljši rezultat pri $k = 1$, kjer so pri vseh treh utežeh odstranjene le povezave med kupci, ki nimajo skupnih izdelkov.

Iz grafa na sliki 6.2 je pri $k = 1$ lepo razviden skupen vrh ocen vseh treh uteži, nakar ocena tretje uteži strmo pade pod 1 %. Povezavo na grafu kupcev tretja utež uteži glede na povprečno število kategorij skupnih izdelkov. Vzrok za strm padec mora biti potemtakem majhen razpon kategorij, v katerih kupujejo posamezni kupci. Slednje lahko potrdimo s statističnim podatkom iz poglavja 5, ki pravi, da je povprečno število kategorij, v katerih kupuje kupec, enako 1,7.



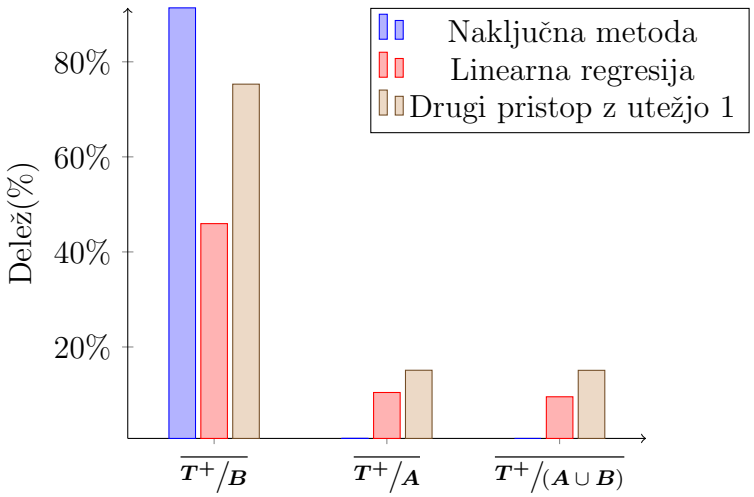
Slika 6.2: Graf ocene $\overline{T^+/(A \cup B)}$ v odvisnosti od k za vse tri uteži pri prvem pristopu

Pristop 2: Vsi sosedje

Drugi pristop, kjer metoda napovedovanja iz grafa kupcev pridobi vse sosede, se je izkazala za mnogo boljšo od prvega pristopa. Ocene rezultatov potrjujejo, da je drugi pristop naše rešitve boljši tudi od regresijskih metod, kar je razvidno iz diagrama na sliki 6.3.

Iz istega razloga kot pri prvem pristopu so tudi tukaj ocene pri vseh treh utežeh (tabela 6.10, tabela 6.11 in tabela 6.12) za vhodni parameter $k = 1$ enake in hkrati najboljše. Tedaj ocena tretjega ocenjevalnega kriterija $\overline{T^+/(A \cup B)}$ z vrednostjo 15 % izboljša najboljši rezultat linearne regresije za faktor 1,5. Napovedna množica se zmanjša s 15 % na 10 %, kar je tudi eden izmed posrednih vzrokov za zmanjšanje nepravilno napovedanih kupcev F^+ in posledično večji dobiček, ki je 11,8.

Iz stolpca $\overline{A/B}$ tabele 6.10 je razvidno, da je delež napovedanih kupcev v primerjavi s prvim pristopom narasel z 92 % na 94 %, kar potrjuje, da izbira pristopa vpliva na napoved kupcev iz polnega grafa.



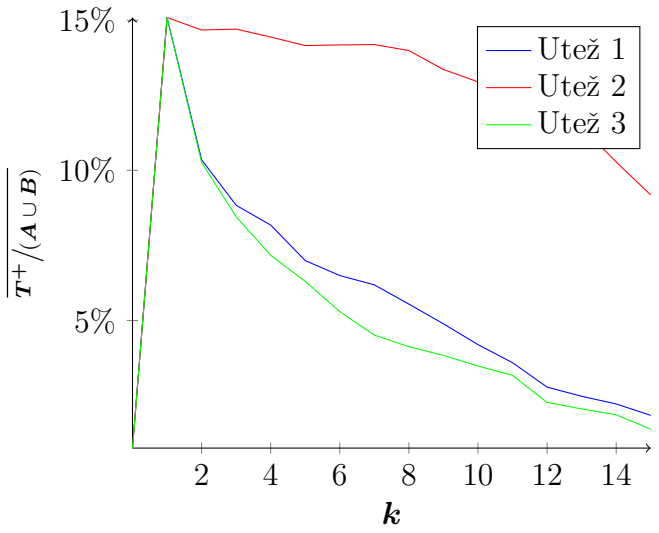
Slika 6.3: Diagram ocen prvih treh ocenjevalnih kriterijev pri naključni metodi, linearni regresiji in drugemu pristopu z utežjo 1

k	$\overline{T^+/B}$	$\overline{T^+/A}$	$\overline{T^+/(A \cup B)}$	Dobiček	$\overline{A/B}$
0	76,05 %	0,75 %	0,75 %	−3,91	94,07 %
1	75,30 %	15,11 %	15,09 %	11,84	10,55 %
2	61,74 %	11,31 %	10,35 %	8,31	9,63 %
14	5,69 %	4,25 %	2,22 %	2,9	1,35 %
15	4,20 %	4,34 %	1,84 %	2,99	1,01 %

Tabela 6.10: Ocene napovedovanja kupcev z drugim pristopom in utežjo 1

k	$\overline{T^+/B}$	$\overline{T^+/A}$	$\overline{T^+/(A \cup B)}$	Dobiček	$\overline{A/B}$
0	76,05 %	0,75 %	0,75 %	−3,91	94,07 %
1	75,30 %	15,11 %	15,09 %	11,84	10,55 %
2	74,44 %	14,69 %	14,67 %	11,47	10,53 %
14	46,60 %	10,99 %	10,28 %	8,47	7,73 %
15	43,64 %	10,09 %	9,18 %	7,57	7,46 %

Tabela 6.11: Ocene napovedovanja kupcev z drugim pristopom utežjo 2



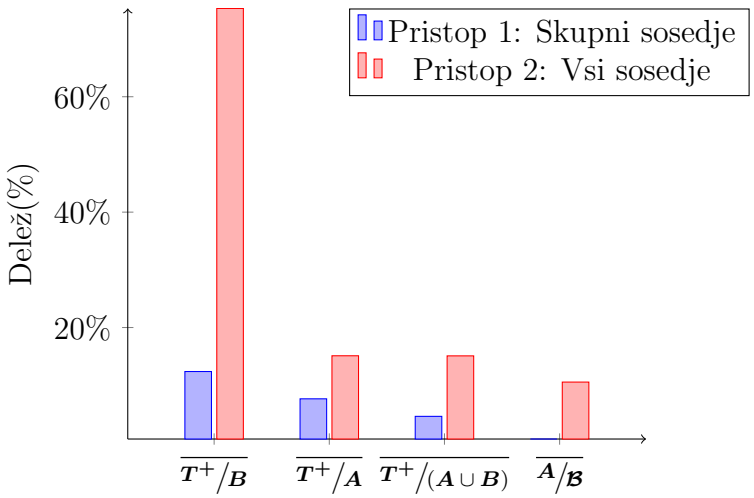
Slika 6.4: Graf ocene $\overline{T^+/(A \cup B)}$ v odvisnosti od k za vse tri uteži pri drugem pristopu

k	$\overline{T^+/B}$	$\overline{T^+/A}$	$\overline{T^+/(A \cup B)}$	Dobiček	$\overline{A/B}$
0	76,05 %	0,75 %	0,75 %	−3,91	94,07 %
1	75,30 %	15,11 %	15,09 %	11,84	10,55 %
2	60,93 %	11,32 %	10,27 %	8,32	9,57 %
14	3,68 %	4,57 %	1,87 %	3,25	0,73 %
15	2,49 %	3,17 %	1,38 %	2,14	0,50 %

Tabela 6.12: Ocene napovedovanja kupcev z drugim pristopom in utežjo 3

S primerjavo vseh treh uteži med seboj, kot je prikazano na grafu v sliki 6.4, opazimo, da ocena tretjega ocenjevalnega kriterija uteži 2 v odvisnosti od k pada počasneje kot ostali dve uteži. Iz tabele 6.11 je razvidno, da tudi velikost napovedne množice pada počasneje kot pri ostalih dveh utežeh. Spomnimo se, da utež 2 vzame tudi količino naročenih izdelkov iz naročil kupca in ne samo števila izdelkov. Potemtakem je počasno padanje smiselno, saj je utež 2 razpon vrednosti uteži povezav na grafu povečala in tako se z enim korakom stopnje omejevanja grafa odstrani manjše število povezav.

Pristop vseh sosedov, ki uporablja utež 3 oziroma utež skupnih kategorij,



Slika 6.5: Diagram ocen prvih treh ocenjevalnih kriterijev in velikost napovedne množice pri prvem in drugem pristopu

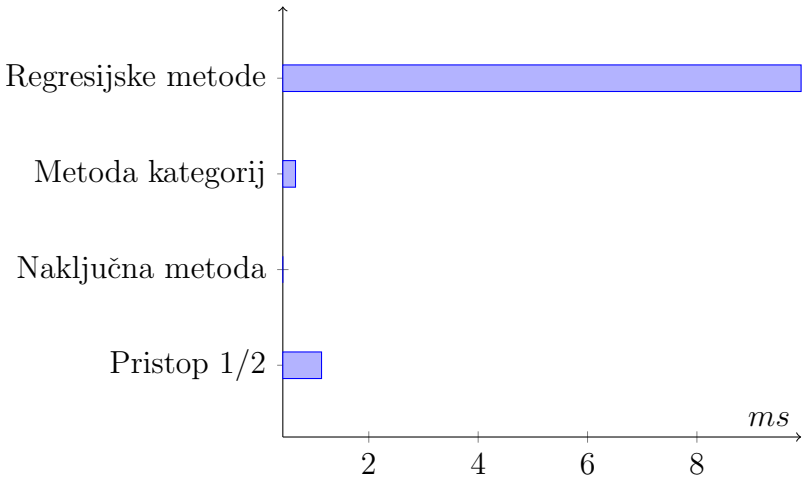
se je izkazal s podobno uspešnostjo kot pristop vseh sosedov z utežjo 1.

Uspešnost drugega pristopa oziroma pristopa vseh sosedov pa je pri uteži 3 nedvomno boljša od pristopa skupnih sosedov. Iz primerjave velikosti napovedne množice $\overline{A/\mathcal{B}}$, prikazane na sliki 6.5, je razvidno, da prvi pristop v primerjavi z drugim pristopom veliko bolj omeji napovedno množico in iz množice izpusti tudi velik delež dejanskih kupcev $\overline{T^+/B}$.

Časi izvajanja pri napovedih kupcev

Poglejmo si čase izvajanja metod, ki smo jih uporabili za napovedovanje kupcev. V stolpčnem diagramu na sliki 6.6 so prikazani časi povprečnega izvajanja napovedi kupcev za 100 izdelkov v milisekundah.

Pričakovano najhitreje sta se izvajali naključna in metoda kategorij. Pristopa, ki smo ju razvili, sta za 100 izdelkov napovedala kupce v povprečno 1,1 milisekunde. Regresijske metode so 100 napovedi izvajale najdlje, in sicer 10 milisekund.



Slika 6.6: Diagram povprečnih časov izvajanja napovedi kupcev za posamezno napoved

6.4 Ovrednotenje napovedovanja izdelkov

Enako kot pri rezultatih napovedovanja kupcev bomo tudi tukaj najprej prikazali ocene oziroma uspešnost primerjalnih metod ter jih na koncu primerjali z našima pristopoma.

Primerjalne metode

V nadaljevanju bodo v tabelah prikazane ocene pri okrnjenem naboru vhodnega parametra $\boldsymbol{k}$. Ocene pri vseh vhodnih parametrih za napovedovanje izdelkov so prikazane v dodatku B.

Naključna metoda, katere ocene so prikazane v tabeli 6.13, se je najbolje odrezala pri vhodnem parametru $\boldsymbol{k} = 93$. Ocena po tretjem ocenjevalnem kriteriju je 0,54 %, kar je pričakovano nizko. Delež odkritih kupcev je 93,26 %, vendar je po deležu drugega ocenjevalnega kriterija $\overline{T^+/C}$ razvidno, da je veliko napovedanih kupcev napovedanih napačno.

k	$\overline{T^+/I}$	$\overline{T^+/C}$	$\overline{T^+/(C \cup I)}$	$\overline{C/\mathcal{I}}$
100	100,00 %	0,54 %	0,54 %	100,00 %
93	93,28 %	0,54 %	0,54 %	92,88 %
86	86,48 %	0,54 %	0,54 %	85,94 %
80	80,08 %	0,54 %	0,54 %	79,86 %
6	6,07 %	0,54 %	0,50 %	5,90 %
0	0,00 %	0,00 %	0,00 %	0,00 %

Tabela 6.13: Ocene napovedovanja izdelkov z naključno metodo

Ocene metode kategorij v tabeli 6.14 prikazujejo, da so napovedi te boljše proti napovedi naključne metode. Tretji ocenjevalni kriterij je metodo kategorij najbolje ocenil pri vhodnem parametru $k = 8$ z oceno 4,25 %. Tedaj je metoda odkrila 15,81 % dejanskih izdelkov, ki sestavljajo 5,37 % napovedne množice.

k	$\overline{T^+/I}$	$\overline{T^+/C}$	$\overline{T^+/(C \cup I)}$	$\overline{C/\mathcal{I}}$
0	0,00 %	0,00 %	0,00 %	0,00 %
1	2,63 %	6,25 %	2,63 %	0,17 %
7	13,23 %	5,21 %	3,97 %	1,22 %
8	15,81 %	5,37 %	4,25 %	1,39 %
9	16,89 %	5,09 %	4,12 %	1,56 %
14	22,99 %	4,58 %	3,98 %	2,43 %
15	24,04 %	4,54 %	3,99 %	2,60 %

Tabela 6.14: Ocene napovedovanja izdelkov z metodo kategorij

Tako kot pri napovedovanju kupcev se je izmed regresijskih metod najbolje odrezala linearna regresija, katere ocene rezultatov so prikazane v tabeli 6.15. Tretji ocenjevalni kriterij $\overline{T^+/(C \cup I)}$ je linearno regresijo najbolje ocenil pri vhodnem parametru $k = 0{,}13$ z oceno 29,8 %. Pri tem je bilo odkritih 73 % dejanskih izdelkov in 33 % napovedne množice je pravilno napovedanih izdelkov.

k	$\overline{T^+/I}$	$\overline{T^+/C}$	$\overline{T^+/(C \cup I)}$	$\overline{C/\mathcal{I}}$
0	92,23 %	1,18 %	1,18 %	53,02 %
0,07	83,20 %	25,70 %	25,00 %	3,87 %
0,13	73,62 %	31,46 %	29,81 %	2,95 %
0,2	64,03 %	35,55 %	32,27 %	2,33 %
0,87	28,88 %	34,82 %	25,31 %	0,89 %
0,93	27,52 %	33,55 %	24,34 %	0,87 %
1	26,23 %	32,25 %	23,41 %	0,86 %

Tabela 6.15: Ocene napovedovanja izdelkov z linearno regresijo

Regresija naključnih gozdov se je po vseh kriterijih odrezala slabše od linearne regresije. Iz tabele 6.16 je razviden najboljši rezultat pri vhodnem parametru $k = 0{,}13$, kjer sta drugi $\overline{T^+/C} = 27{,}7\%$ in tretji ocenjevalni kriterij $\overline{T^+/(C \cup I)} = 23{,}8\%$ RF-regresijo ocenila najbolje. Največji delež izdelkov, 63,7 %, je RF-regresija odkrila pri vhodnem parametru $k = 0{,}07$.

k	$\overline{T^+/I}$	$\overline{T^+/C}$	$\overline{T^+/(C \cup I)}$	$\overline{C/\mathcal{I}}$
0	100,00 %	0,54 %	0,54 %	100,00 %
0,07	63,91 %	23,87 %	21,43 %	3,87 %
0,13	53,85 %	27,67 %	23,72 %	3,18 %
0,2	53,85 %	27,67 %	23,72 %	3,18 %
0,87	14,13 %	19,24 %	10,46 %	1,60 %
0,93	12,06 %	16,40 %	8,85 %	1,51 %
1	12,06 %	16,40 %	8,85 %	1,50 %

Tabela 6.16: Ocene napovedovanja izdelkov z RF-regresijo

Pri regresiji k_{knn} skupnih sosedov oziroma KNN-regresiji ocene kažejo na slabšo uspešnost v primerjavi z linearno ali RF-regresijo. Vendar se je KNN-regresija po oceni $\overline{T^+/(C \cup I)}$ odrezala bolje od metode kategorij. Iz tabele 6.17 je razviden najboljši rezultat pri vhodnem parametru $k = 0{,}07$.

k	$\overline{T^+/I}$	$\overline{T^+/C}$	$\overline{T^+/(C \cup I)}$	$\overline{C/\mathcal{I}}$
0	100,00 %	0,54 %	0,54 %	100,00 %
0,07	9,54 %	11,58 %	5,50 %	2,26 %
0,13	9,54 %	11,58 %	5,50 %	2,26 %
0,2	9,54 %	11,58 %	5,50 %	2,26 %
0,87	3,55 %	4,14 %	1,75 %	1,48 %
0,93	3,55 %	4,14 %	1,75 %	1,48 %
1	3,55 %	4,14 %	1,75 %	1,48 %

Tabela 6.17: Ocene napovedovanja izdelkov s KNN-regresijo ($k_{knn} = 3$)

Logistična regresija, katere ocene so prikazane v tabeli 6.18, je po tretjem ocenjevalnem kriteriju $\overline{T^+/(C \cup I)}$ za las premagala metodo kategorij. Z napovedjo je slednja odkrila 5,35 % dejanskih kupcev, ki tvorijo 10,8 % celotne napovedne množice C.

$\overline{T^+/I}$	$\overline{T^+/C}$	$\overline{T^+/(C \cup I)}$	$\overline{A/\mathcal{B}}$
5,35 %	10,71 %	4,88 %	0,09 %

Tabela 6.18: Ocene napovedovanja izdelkov z logistično regresijo

Pristop 1: Skupni sosedje

Poglejmo si ocene rezultatov drugega pristopa za napovedovanje izdelkov. Rezultate prvega kot tudi drugega pristopa smo testirali pri prvih treh utežeh, definiranih v poglavju 4:

- utež 1 ... število skupnih kupcev,
- utež 2 ... število skupnih kupcev z naročenimi količinami,
- utež 3 ... ali sta izdelka v skupni kategoriji.

Najboljši rezultat je po oceni tretjega ocenjevalnega kriterija prvi pristop z utežjo 1 napovedal pri vhodnem parametru $k = 8$. Tedaj je v povprečju

9,4 % napovedanih izdelkov odkrilo 39,8 % dejanskih izdelkov, ki jih je kupec kupoval v prihodnosti.

k	$\overline{T^+/I}$	$\overline{T^+/C}$	$\overline{T^+/(C \cup I)}$	$\overline{C/\mathcal{I}}$
0	47,16 %	0,33 %	0,33 %	94,41 %
7	40,60 %	9,21 %	8,28 %	4,11 %
8	39,85 %	9,42 %	8,38 %	3,93 %
9	38,64 %	9,48 %	8,36 %	3,75 %
10	37,76 %	9,40 %	8,27 %	3,63 %
36	24,47 %	9,17 %	7,42 %	1,75 %

Tabela 6.19: Ocene napovedovanja izdelkov s prvim pristopom in utežjo 1

Pri drugi uteži so ocene (tabela 6.20) z manjšimi razlikami med različnimi vhodnimi parametri k. To pomeni, da je nabor vrednosti uteži povezav tukaj večji, kar je posledica funkcije uteži 2, ki upošteva tudi količine naročenih izdelkov. Vrh ocene po tretjem ocenjevalnem kriteriju je okoli vhodnega parametra $k = 10$, kjer algoritem skupnih sosedov odkrije skoraj 40 % vseh dejanskih izdelkov in je ocena tretjega ocenjevalnega kriterija enaka 8,5 %.

k	$\overline{T^+/I}$	$\overline{T^+/C}$	$\overline{T^+/(C \cup I)}$	$\overline{C/\mathcal{I}}$
0	47,16 %	0,33 %	0,33 %	94,41 %
9	40,34 %	9,42 %	8,44 %	4,21 %
10	39,80 %	9,48 %	8,45 %	4,11 %
11	39,30 %	9,52 %	8,45 %	4,03 %
12	38,53 %	9,57 %	8,44 %	3,91 %
36	26,64 %	9,24 %	7,38 %	2,41 %

Tabela 6.20: Ocene napovedovanja izdelkov s prvim pristopom in utežjo 2

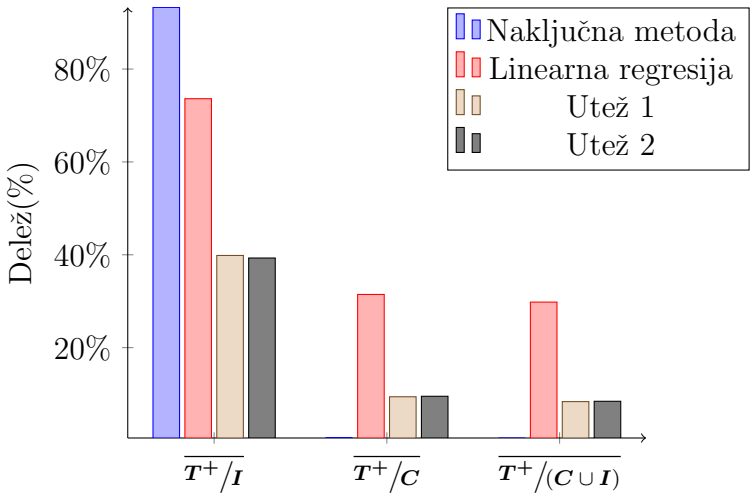
Funkcija tretje uteži pri napovedovanju kupcev ima logičen $\{0, 1\}$ nabor vrednosti. Prav zato omejevanje grafa s stopnjo omejevanja k nad 1 ni smiselno, saj graf izdelkov nima povezav, večjih od 1. Ocene rezultatov

tretje uteži, ki so prikazane v tabeli 6.21, niso primerljive z nobeno od boljših metod, saj ne izboljšajo rezultatov naključne metode.

k	$\overline{T^+/I}$	$\overline{T^+/C}$	$\overline{T^+/(C \cup I)}$	$\overline{C/\mathcal{I}}$
0	47,16 %	0,33 %	0,33 %	94,41 %
1	24,14 %	0,53 %	0,53 %	23,45 %

Tabela 6.21: Ocene napovedovanja izdelkov s prvim pristopom in utežjo 3

Pristop skupnih sosedov se je v povprečju odrezal bolje od obeh enostavnih metod (naključna metoda in metoda kategorij), vendar so bile regresijske metode pri napovedovanju izdelkov boljše. Najbolje od vseh metod se je odrezala linearna regresija, ki je najboljši primer drugega pristopa po oceni tretjega ocenjevalnega kriterija presegla za faktor 3,5 (slika 6.7).



Slika 6.7: Diagram ocen prvih treh ocenjevalnih kriterijev pri napovedovanju izdelkov

k	$\overline{T^+/I}$	$\overline{T^+/C}$	$\overline{T^+/(C \cup I)}$	$\overline{C/\mathcal{I}}$
0	87,13 %	0,53 %	0,53 %	95,81 %
1	84,89 %	1,45 %	1,44 %	40,70 %

Tabela 6.22: Ocene napovedovanja izdelkov z drugim pristopom in utežjo 3

Pristop 2: Vsi sosedje

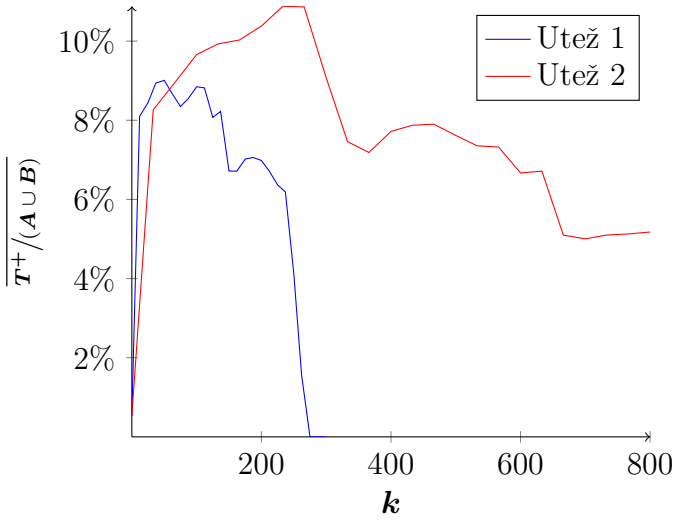
Drugi pristop oziroma pristop vseh sosedov se je tako kot pri prvem pristopu najslabše odrezal z uporabo tretje uteži (tabela 6.22). Klub temu pa je bil boljši od prvega pristopa z utežjo 3.

V drugem pristopu, kjer so izdelki napovedi izbrani s pomočjo skupnih sosedov, se je okno rezultatov glede na parameter k v primerjavi z napovedovanjem kupcev zelo raztegnilo. To je posledica velikega razpona nabora vrednosti in razpršenosti uteži povezav v grafu med izdelki, kar sledi iz dejstva, da posamezen izdelek lahko nastopa v veliko več naročilih kot posamezen kupec. To potrjujeta povprečni stopnji vozlišč, ki smo jih izračunali za vsako stran dvodelnega grafa $G_{\mathcal{BI}}$. Povprečna stopnja vozlišča kupca je enaka 10, medtem ko je povprečna stopnja vozlišča izdelka enaka 72.

k	$\overline{T^+/I}$	$\overline{T^+/C}$	$\overline{T^+/(C \cup I)}$	$\overline{C/\mathcal{I}}$
0	87,13 %	0,53 %	0,53 %	95,81 %
33	82,18 %	8,35 %	8,26 %	6,85 %
133	68,59 %	10,40 %	9,93 %	4,39 %
166	64,31 %	11,03 %	10,03 %	3,92 %
200	60,03 %	11,54 %	10,38 %	3,38 %
233	55,14 %	12,27 %	10,88 %	2,82 %
266	45,52 %	12,73 %	10,86 %	2,11 %
800	12,61 %	8,03 %	5,18 %	0,43 %

Tabela 6.23: Ocene napovedovanja izdelkov z drugim pristopom in utežjo 2

Ocene rezultatov pri uteži 1 in uteži 2 so do omejevanja grafa s stopnjo omejevanja k približno 70 videti podobne, vendar se kasneje utež 2 izkaže za boljšo, kar je lepo razvidno iz grafa na sliki 6.8. Utež 2 doseže vrh pri omejevanju grafa okoli 233, kjer po tretjem ocenjevalnem kriteriju dobi oceno 10,9 %. Pri tem odkrije približno 55 % dejanskih izdelkov, ki sestavljajo približno 12,3 % napovedne množice. Nekaj dobrih ocen pri ostalih vrednostih k za drugo utež je prikazanih v tabeli 6.23.



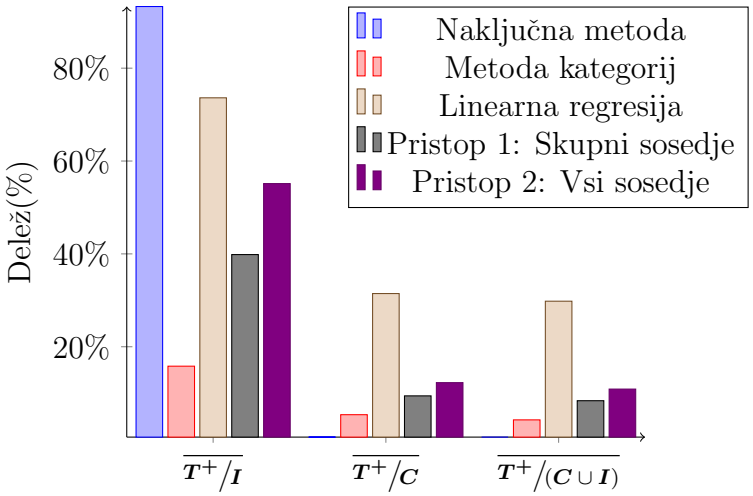
Slika 6.8: Graf ocene $\overline{T^+/(C \cup I)}$ v odvisnosti od k za prvo in drugo utež pri drugem pristopu napovedovanja izdelkov

k	$\overline{T^+/I}$	$\overline{T^+/C}$	$\overline{T^+/(C \cup I)}$	$\overline{C/\mathcal{I}}$
0	87,13 %	0,53 %	0,53 %	95,81 %
25	74,93 %	8,69 %	8,45 %	6,02 %
37	70,50 %	9,35 %	8,94 %	5,34 %
50	66,35 %	9,42 %	9,01 %	4,91 %
62	62,23 %	9,78 %	8,68 %	4,56 %
262	1,88 %	4,25 %	1,57 %	0,10 %

Tabela 6.24: Ocene napovedovanja izdelkov z drugim pristopom in utežjo 1

Drugi pristop z utežjo 1 po tretjem ocenjevalnem kriteriju $\overline{T^+/(C \cup I)}$ najbolje napove izdelke pri $k = 50$, kjer je ocena enaka 9 % (tabela 6.24). Pri $k = 50$ napovedna množica C vsebuje 9,4 % pravilno napovedanih izdelkov, ki odkrijejo 66,4 % dejanskih izdelkov, ki jih je kupec kupoval v prihodnosti.

Enako kot pri napovedovanju kupcev je tudi pri napovedovanju izdelkov od naše rešitve boljši drugi pristop, pristop vseh sosedov. Z drugim pristopom smo rezultate prvega pristopa izboljšali za faktor 1,2, vendar je linearna regresija izdelke napovedala bolj natančno kot oba naša pristopa, kar je razvidno iz stolpčnega diagrama na sliki 6.9.

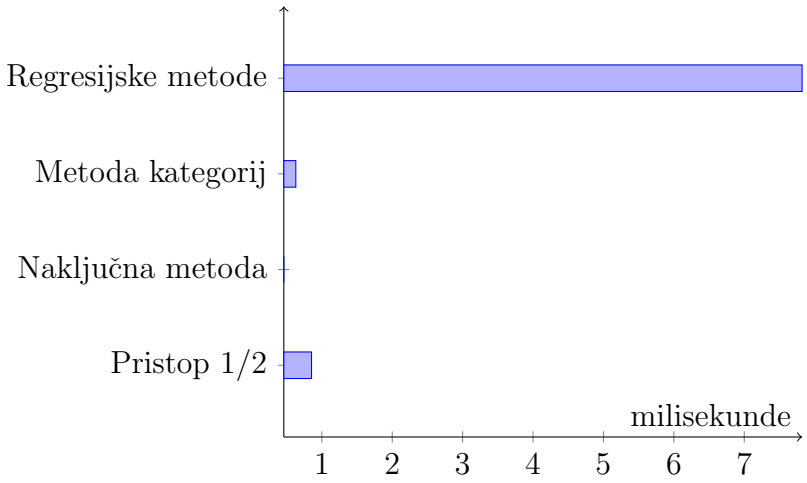


Slika 6.9: Diagram ocen prvih treh ocenjevalnih kriterijev pri napovedovanju izdelkov

Časi izvajanja pri napovedih izdelkov

V stolpčnem diagramu na sliki 6.10 so prikazani časi povprečnega izvajanja napovedi izdelkov za 100 kupcev v milisekundah.

Najhitrejši metodi, metoda kategorij in naključna metoda, sta za 100 kupcev izdelke napovedali v 0,5 milisekunde. Najpočasnejše so bile regresijske metode, ki so se v povprečju 100 napovedi izvajale 8 milisekund. Pristop 1 in 2 sta za 100 napovedi porabila v povprečju 0,9 milisekunde.



Slika 6.10: Diagram časov izvajanja za posamezno napoved pri napovedovanju izdelkov

6.5 Ovrednotenje napovedovanja s hibridno metodo

Pri hibridni metodi smo združili metodi napovedovanja kupcev z drugim pristopom in prvo utežjo ter napovedovanje izdelkov prav tako z drugim pristopom, vendar z drugo utežjo. Slednja konfiguracija je bila izbrana na podlagi rezultatov posameznih pristopov in uteži ter testiranja. V hibridni metodi nastopata dve omejevanji grafa in potemtakem dve stopnji omejevanja: prva stopnja omejevanja $\boldsymbol{k}_1$ za omejevanje grafa kupcev pri napovedovanju kupcev, druga stopnja $\boldsymbol{k}_2$ pa za omejevanje grafa izdelkov.

Ocene rezultatov prvega pristopa hibridne metode, kjer se napovedi kupcev in izdelkov potrjujejo, so prikazane v priponki A s povzetkom v tabeli 6.25. Tukaj hibridna metoda doseže vrh pri $\boldsymbol{k}_1 = 1$ in $\boldsymbol{k}_2 = 1$. Tedaj po tretjem ocenjevalnem kriteriju dobi oceno 10,9 % in dobiček je ocenjen z 8,62 na napovedanega kupca. Prvi pristop ne izboljša dosedanjega najboljšega rezultata napovedovanja kupca, ki ga ima pristop vseh sosedov z utežjo 1, katerega ocena po tretjem ocenjevalnem kriteriju znaša 15,1 %.

k_1	k_2	$\overline{T^+/B}$	$\overline{T^+/A}$	$\overline{T^+/(A \cup B)}$	Dobiček	$\overline{A/\mathcal{B}}$
0	0	67,61 %	0,73 %	0,72 %	−3,94	94,04 %
1	0	66,86 %	11,93 %	10,91 %	8,62	10,52 %
1	1	66,86 %	11,93 %	10,91 %	8,62	10,52 %
2	1	66,00 %	11,52 %	10,49 %	8,24	10,50 %
0	2	60,85 %	9,45 %	8,86 %	6,58	10,38 %
1	2	60,85 %	9,45 %	8,86 %	6,58	10,38 %

Tabela 6.25: Ocene napovedovanja kupcev hibridne metode s prvim pristopom

Ocene rezultatov drugega pristopa, kjer se napovedi dopolnjujejo, so prikazane v tabeli 6.26. Pri $k_1 = 1$ in $k_2 = 1$ tabela prikazuje izboljšanje dosedanjega najboljšega rezultata s 15,1 % na 22,3 %. Pri najbolje ocenjenem rezultatu je metoda dobiček povišala na 19,2 in je z 22,4 % pravilno napovedanimi kupci v napovedni množici odkrila 86 % vseh dejanskih kupcev. S temi ocenami rezultatov je drugi pristop hibridne metode pri testih najbolje napovedoval kupce.

k_1	k_2	$\overline{T^+/B}$	$\overline{T^+/A}$	$\overline{T^+/(A \cup B)}$	Dobiček	$\overline{A/\mathcal{B}}$
0	0	86,79 %	0,76 %	0,76 %	−3,91	94,07 %
1	0	86,79 %	0,76 %	0,76 %	−3,91	94,07 %
1	1	86,04 %	22,39 %	22,25 %	19,17	10,55 %
2	1	86,04 %	22,39 %	22,25 %	19,17	10,55 %
0	2	76,05 %	0,75 %	0,75 %	−3,91	94,07 %
1	2	75,30 %	15,11 %	15,09 %	11,84	10,55 %

Tabela 6.26: Ocene napovedovanja kupcev hibridne metode z drugim pristopom

Poglavje 7

Sklepne ugotovitve

V sklopu magistrskega dela smo definirali model napovedovanja, ki opisuje problematiko napovedovanja izdelkov in kupcev. Ob modelu so bili definirani tudi ocenjevalni kriteriji, ki na podlagi napovedne in testne množice ocenjujejo napovedi iz različnih pogledov (delež pravilno napovedanih, dobiček ...). Za rešitev problema napovedovanja smo s teorijo grafov in množic razvili pristope, ki preko dveh tipov grafov napovedujejo kupce in izdelke. Pri tem razvita rešitev omogoča napovedovanje za različne entitete (kupce ali izdelke) brez ponovne gradnje vseh grafov oziroma struktur pristopa.

Natančnost napovedovanja razvitih pristopov smo testirali in primerjali s šestimi primerjalnimi metodami: naključna metoda, metoda kategorij, linearna regresija, logistična regresija, regresija naključnih gozdov in regresija k-najbližjih sosedov. Za smiselno ovrednotenje in testiranje smo obdelali in pripravili podatke o resničnih prodajah spletnega prodajnega mesta podjetja Amazon. Po ocenah in ovrednotenju rezultatov testiranja sta drugi pristop in hibridna metoda kupce napovedala bolje od vseh primerjalnih metod, tudi od regresijskih metod, ki so ene od najpogosteje uporabljenih metod v napovedni analitiki. Spletnim prodajalcem s tem omogočamo optimizacijo izbire kupcev pri promocijah in drugih aktivnostih, kjer prodajalci uporabljajo napovedno analitiko.

Rezultati nakazujejo, da alternativnih deterministčnih pristopov, kot je

pristop s teorijo grafov, ne gre vedno zavreči v zamenjavo za metode regresije in podatkovno rudarjenje. Prednost deterministčnih metod je tudi hitrost izvajanja, saj so v večini primerov hitrejše od regresijskih metod ali drugih metod podatkovnega rudarjenja.

Odprto področje napovedovanja s pomočjo pristopa teorije grafov smo v sklopu tega magistrskega dela pripeljali do točke, kjer je morebitno nadaljnje delo in dodaten razvoj predlaganih pristopov utemeljeno. Dodatno motivacijo predstavlja uspešnost napovedovanja hibridne metode, ki je nadgradnja grafovskega modela napovedovanja. Pristop napovedovanja kupcev bi lahko prilagodili za pospeševanje prodaje novih izdelkov, ki jih prodajalci želijo postaviti na trg spletne prodaje. Nadaljnje delo bi lahko vključevalo tudi posplošitev grafovskega modela napovedovanja na model odločanja ali združitev predlaganih pristopov z regresijskimi metodami na način, kjer bi slednje uteževale značilke, uporabljene na utežeh v grafih. Neverjetna rast spletne prodaje prodajalcem prinaša dodatne izzive s čedalje večjo količino podatkov, zato je smiselno razmišljati o razvoju predlaganih pristopov v domeni analize velikih podatkov (angl. big data analysis).

Dodatek A

Rezultati napovedovanja kupcev

A.1 Primerjalne metode

Naključna metoda

k	$\overline{T^+/B}$	$\overline{T^+/A}$	$\overline{T^+/(A \cup B)}$	Dobiček	$\overline{A/B}$
100	100,00 %	0,76 %	0,76 %	−4,2	100,00 %
93	91,35 %	0,76 %	0,76 %	−4,2	93,00 %
86	86,76 %	0,76 %	0,76 %	−4,2	86,00 %
80	79,39 %	0,76 %	0,76 %	−4,2	79,99 %
73	72,25 %	0,76 %	0,75 %	−4,2	72,99 %
66	67,17 %	0,75 %	0,74 %	−4,21	65,99 %
60	58,52 %	0,76 %	0,74 %	−4,2	60,00 %
53	53,50 %	0,76 %	0,74 %	−4,2	52,99 %
46	44,87 %	0,76 %	0,73 %	−4,2	45,99 %
40	41,78 %	0,78 %	0,74 %	−4,18	39,98 %
33	33,80 %	0,78 %	0,72 %	−4,18	32,98 %
26	25,49 %	0,75 %	0,69 %	−4,21	25,98 %
20	19,41 %	0,76 %	0,67 %	−4,2	19,99 %
13	13,57 %	0,75 %	0,62 %	−4,21	12,99 %
6	4,64 %	0,74 %	0,50 %	−4,22	5,99 %
0	0,00 %	0,00 %	0,00 %	0	0,00 %

Metoda kategorij

k	$\overline{T^+/B}$	$\overline{T^+/A}$	$\overline{T^+/(A \cup B)}$	Dobiček	$\overline{A/\mathcal{B}}$
0	94,79 %	4,81 %	4,76 %	0,06	42,64 %
1	91,50 %	5,73 %	5,57 %	1,03	37,00 %
2	90,50 %	6,18 %	5,91 %	1,55	36,11 %
3	89,22 %	6,64 %	6,27 %	2,03	35,43 %
4	87,69 %	6,73 %	6,31 %	2,15	34,95 %
5	86,06 %	6,73 %	6,30 %	2,14	34,57 %
6	84,57 %	6,75 %	6,30 %	2,16	34,07 %
7	82,82 %	6,88 %	6,43 %	2,3	33,41 %
8	80,17 %	6,42 %	5,89 %	1,86	32,41 %
9	77,17 %	6,37 %	5,77 %	1,81	30,81 %
10	66,74 %	6,17 %	5,53 %	1,61	25,15 %
11	54,21 %	5,70 %	4,99 %	1,15	19,72 %
12	40,59 %	5,55 %	4,74 %	1,01	14,20 %
13	36,35 %	5,66 %	4,77 %	1,13	11,73 %
14	31,33 %	5,27 %	4,32 %	0,74	9,72 %
15	26,07 %	5,14 %	4,28 %	0,6	7,88 %

Linearna regresija

k	$\overline{T^+/B}$	$\overline{T^+/A}$	$\overline{T^+/(A \cup B)}$	Dobiček	$\overline{A/\mathcal{B}}$
0	85,16 %	1,01 %	0,99 %	−3,94	80,60 %
0,07	45,95 %	10,42 %	9,52 %	6,38	14,69 %
0,13	35,56 %	10,54 %	9,36 %	6,61	8,57 %
0,2	29,07 %	10,61 %	9,22 %	6,71	5,80 %
0,27	23,10 %	9,18 %	7,66 %	5,3	4,74 %
0,33	19,38 %	8,34 %	6,75 %	4,49	3,82 %
0,4	16,56 %	7,60 %	6,04 %	3,76	3,08 %
0,47	15,01 %	7,40 %	5,82 %	3,58	2,68 %
0,53	13,36 %	6,78 %	5,05 %	2,99	2,40 %
0,6	12,15 %	6,71 %	4,93 %	2,91	2,13 %
0,67	11,18 %	6,16 %	4,71 %	2,38	1,89 %
0,73	10,60 %	6,04 %	4,59 %	2,28	1,67 %
0,8	9,54 %	5,28 %	4,13 %	1,52	1,49 %
0,87	9,30 %	5,26 %	4,09 %	1,51	1,37 %
0,93	8,36 %	5,11 %	3,94 %	1,36	1,28 %
1	7,87 %	4,69 %	3,65 %	0,96	1,24 %

KNN regresija

k	$\overline{T^+/B}$	$\overline{T^+/A}$	$\overline{T^+/(A \cup B)}$	Dobiček	$\overline{A/\mathcal{B}}$
0	100,00 %	0,76 %	0,76 %	−4,2	100,00 %
0,07	75,10 %	3,34 %	3,17 %	−0,69	50,60 %
0,13	70,67 %	3,30 %	3,08 %	−0,56	49,11 %
0,2	70,67 %	3,30 %	3,08 %	−0,56	49,11 %
0,27	68,85 %	3,86 %	3,35 %	0,1	47,82 %
0,33	65,88 %	4,04 %	3,46 %	0,36	46,80 %
0,4	65,88 %	4,04 %	3,46 %	0,36	46,80 %
0,47	62,41 %	4,01 %	3,19 %	0,42	45,83 %
0,53	59,34 %	3,54 %	2,67 %	0,05	45,14 %
0,6	59,34 %	3,54 %	2,67 %	0,05	45,14 %
0,67	57,01 %	3,05 %	2,44 %	−0,34	44,39 %
0,73	54,19 %	2,36 %	2,07 %	−0,99	43,74 %
0,8	54,19 %	2,36 %	2,07 %	−0,99	43,74 %
0,87	51,11 %	2,54 %	2,22 %	−0,71	42,29 %
0,93	48,92 %	2,80 %	2,25 %	−0,28	41,15 %
1	48,92 %	2,80 %	2,25 %	−0,28	41,15 %

Regresija naključnih gozdov

k	$\overline{T^+/B}$	$\overline{T^+/A}$	$\overline{T^+/(A \cup B)}$	Dobiček	$\overline{A/\mathcal{B}}$
0	100,00 %	0,76 %	0,76 %	−4,2	100,00 %
0,07	54,38 %	9,84 %	8,08 %	6,16	15,85 %
0,13	49,63 %	10,40 %	7,84 %	6,86	15,26 %
0,2	49,63 %	10,40 %	7,84 %	6,86	15,26 %
0,27	46,41 %	10,22 %	7,29 %	6,82	14,97 %
0,33	43,40 %	9,58 %	6,33 %	6,3	14,78 %
0,4	43,36 %	9,33 %	6,29 %	6,05	14,78 %
0,47	41,41 %	9,00 %	5,75 %	5,78	14,63 %
0,53	39,80 %	8,61 %	5,26 %	5,5	14,53 %
0,6	39,80 %	8,61 %	5,26 %	5,5	14,53 %
0,67	38,03 %	7,83 %	4,40 %	4,88	14,45 %
0,73	37,53 %	7,87 %	4,27 %	5,04	14,39 %
0,8	37,53 %	7,87 %	4,27 %	5,04	14,39 %
0,87	36,99 %	7,63 %	4,11 %	4,88	14,35 %
0,93	36,45 %	7,41 %	3,91 %	4,75	14,23 %
1	36,45 %	7,42 %	3,92 %	4,75	14,21 %

A.2 Pristop 1: Skupni sosedje

Utež 1

k	$\overline{T^+/B}$	$\overline{T^+/A}$	$\overline{T^+/(A \cup B)}$	Dobiček	$\overline{A/\not B}$
0	37,58 %	0,48 %	0,47 %	−4,2	92,09 %
1	12,38 %	7,64 %	4,61 %	5,83	0,68 %
2	6,16 %	4,91 %	2,80 %	3,81	0,23 %
3	3,97 %	3,25 %	2,23 %	2,58	0,08 %
4	3,59 %	3,18 %	2,10 %	2,71	0,04 %
5	1,92 %	2,40 %	1,45 %	2,11	0,05 %
6	0,85 %	1,93 %	0,72 %	1,66	0,02 %
7	0,19 %	0,42 %	0,13 %	0,24	0,02 %
8	0,14 %	0,42 %	0,12 %	0,26	0,01 %
9	0,09 %	0,66 %	0,09 %	0,51	0,01 %
10	0,11 %	0,49 %	0,10 %	0,38	0,00 %
11	0,04 %	0,37 %	0,04 %	0,22	0,00 %
12	0,02 %	0,74 %	0,02 %	0,56	0,00 %
13	0,05 %	0,62 %	0,04 %	0,5	0,01 %
14	0,06 %	0,49 %	0,06 %	0,38	0,01 %
15	0,00 %	0,08 %	0,00 %	−0,05	0,01 %

Utež 2

k	$\overline{T^+/B}$	$\overline{T^+/A}$	$\overline{T^+/(A \cup B)}$	Dobiček	$\overline{A/\not B}$
0	37,58 %	0,48 %	0,47 %	−4,2	92,09 %
1	12,38 %	7,64 %	4,61 %	5,83	0,68 %
2	8,40 %	6,02 %	3,41 %	4,44	0,35 %
3	6,19 %	5,39 %	2,94 %	4,02	0,18 %
4	4,60 %	4,21 %	2,20 %	3,2	0,11 %
5	3,19 %	4,13 %	2,03 %	3,27	0,08 %
6	1,82 %	2,63 %	1,33 %	1,8	0,06 %
7	0,83 %	1,27 %	0,45 %	0,49	0,05 %
8	0,63 %	1,09 %	0,34 %	0,42	0,04 %
9	0,61 %	1,37 %	0,38 %	0,77	0,03 %
10	0,52 %	0,85 %	0,32 %	0,33	0,02 %
11	0,52 %	0,88 %	0,34 %	0,45	0,02 %
12	0,59 %	1,11 %	0,37 %	0,68	0,02 %
13	0,67 %	1,54 %	0,47 %	1,18	0,02 %
14	0,68 %	1,44 %	0,51 %	1,13	0,02 %
15	0,49 %	0,90 %	0,34 %	0,56	0,02 %

Utež 3

k	$\overline{T^+/B}$	$\overline{T^+/A}$	$\overline{T^+/(A \cup B)}$	Dobiček	$\overline{A/\mathcal{B}}$
0	37,58 %	0,48 %	0,47 %	−4,2	92,09 %
1	12,38 %	7,64 %	4,61 %	5,83	0,68 %
2	2,23 %	3,26 %	0,82 %	2,51	0,15 %
3	0,58 %	1,61 %	0,41 %	1,11	0,04 %
4	0,34 %	1,43 %	0,34 %	1,1	0,03 %
5	0,76 %	1,56 %	0,64 %	1,37	0,02 %
6	0,45 %	1,23 %	0,45 %	1,09	0,01 %
7	0,04 %	0,25 %	0,04 %	0,16	0,01 %
8	0,00 %	0,00 %	0,00 %	−0,1	0,00 %
9	0,00 %	0,00 %	0,00 %	−0,12	0,00 %
10	0,08 %	0,49 %	0,08 %	0,36	0,00 %
11	0,11 %	1,11 %	0,11 %	0,91	0,00 %
12	0,17 %	1,48 %	0,17 %	1,27	0,01 %
13	0,17 %	1,48 %	0,17 %	1,32	0,01 %
14	0,16 %	0,99 %	0,16 %	0,85	0,01 %
15	0,01 %	0,25 %	0,01 %	0,22	0,01 %

A.3 Pristop 2: Vsi sosedje

Utež 1

k	$\overline{T^+/B}$	$\overline{T^+/A}$	$\overline{T^+/(A \cup B)}$	Dobiček	$\overline{A/\mathcal{B}}$
0	76,05 %	0,75 %	0,75 %	−3,91	94,07 %
1	75,30 %	15,11 %	15,09 %	11,84	10,55 %
2	61,74 %	11,31 %	10,35 %	8,31	9,63 %
3	53,68 %	9,74 %	8,84 %	6,89	8,78 %
4	47,68 %	9,92 %	8,18 %	7,25	7,98 %
5	41,19 %	8,65 %	7,00 %	6,14	7,27 %
6	36,61 %	8,29 %	6,50 %	5,88	6,60 %
7	33,01 %	8,54 %	6,19 %	6,18	6,04 %
8	28,54 %	8,02 %	5,55 %	5,71	5,45 %
9	24,10 %	8,60 %	4,89 %	6,32	4,78 %
10	18,75 %	8,43 %	4,20 %	6,32	3,96 %
11	14,22 %	8,86 %	3,60 %	6,79	3,04 %
12	10,00 %	7,41 %	2,78 %	5,65	2,31 %
13	7,46 %	4,80 %	2,48 %	3,28	1,73 %
14	5,69 %	4,25 %	2,22 %	2,9	1,35 %
15	4,20 %	4,34 %	1,84 %	2,99	1,01 %

Utež 2

k	$\overline{T^+/B}$	$\overline{T^+/A}$	$\overline{T^+/(A \cup B)}$	Dobiček	$\overline{A/\mathcal{B}}$
0	76,05 %	0,75 %	0,75 %	−3,91	94,07 %
1	75,30 %	15,11 %	15,09 %	11,84	10,55 %
2	74,44 %	14,69 %	14,67 %	11,47	10,53 %
3	73,86 %	14,75 %	14,70 %	11,56	10,50 %
4	72,55 %	14,48 %	14,44 %	11,33	10,47 %
5	70,74 %	14,22 %	14,15 %	11,06	10,43 %
6	70,00 %	14,29 %	14,18 %	11,13	10,37 %
7	68,82 %	14,47 %	14,19 %	11,36	10,19 %
8	65,68 %	14,50 %	13,99 %	11,45	9,93 %
9	62,09 %	14,32 %	13,36 %	11,37	9,59 %
10	59,01 %	13,82 %	12,95 %	11,06	9,21 %
11	55,86 %	13,67 %	12,83 %	10,94	8,80 %
12	53,37 %	13,08 %	12,29 %	10,36	8,42 %
13	50,10 %	12,32 %	11,46 %	9,72	8,08 %
14	46,60 %	10,99 %	10,28 %	8,47	7,73 %
15	43,64 %	10,09 %	9,18 %	7,57	7,46 %

Utež 3

k	$\overline{T^+/B}$	$\overline{T^+/A}$	$\overline{T^+/(A \cup B)}$	Dobiček	$\overline{A/\mathcal{B}}$
0	76,05 %	0,75 %	0,75 %	−3,91	94,07 %
1	75,30 %	15,11 %	15,09 %	11,84	10,55 %
2	60,93 %	11,32 %	10,27 %	8,32	9,57 %
3	50,35 %	9,90 %	8,46 %	7,06	8,63 %
4	41,67 %	9,01 %	7,17 %	6,38	7,67 %
5	33,70 %	8,98 %	6,30 %	6,51	6,76 %
6	27,84 %	8,13 %	5,30 %	5,76	5,89 %
7	23,08 %	7,23 %	4,51 %	5,01	5,05 %
8	19,33 %	6,45 %	4,13 %	4,4	4,16 %
9	15,94 %	7,10 %	3,84 %	5,13	3,33 %
10	12,63 %	7,70 %	3,49 %	5,77	2,62 %
11	9,54 %	8,61 %	3,17 %	6,74	1,88 %
12	6,28 %	6,60 %	2,27 %	4,98	1,37 %
13	4,71 %	5,19 %	2,06 %	3,7	0,99 %
14	3,68 %	4,57 %	1,87 %	3,25	0,73 %
15	2,49 %	3,17 %	1,38 %	2,14	0,50 %

A.4 Hibridna metoda

Prvi pristop

k1	k2	$\overline{T^+/B}$	$\overline{T^+/A}$	$\overline{T^+/(A \cup B)}$	Dobiček	$\overline{A/B}$
0	0	67,61 %	0,73 %	0,72 %	−3,94	94,04 %
1	0	66,86 %	11,93 %	10,91 %	8,62	10,52 %
2	0	66,00 %	11,52 %	10,49 %	8,24	10,50 %
3	0	65,42 %	11,57 %	10,51 %	8,33	10,47 %
4	0	64,11 %	11,29 %	10,23 %	8,08	10,44 %
5	0	62,30 %	10,98 %	9,89 %	7,77	10,40 %
6	0	61,56 %	11,04 %	9,89 %	7,83	10,34 %
7	0	60,87 %	11,24 %	9,95 %	8,08	10,16 %
0	1	66,86 %	11,93 %	10,91 %	8,62	10,52 %
1	1	66,86 %	11,93 %	10,91 %	8,62	10,52 %
2	1	66,00 %	11,52 %	10,49 %	8,24	10,50 %
3	1	65,42 %	11,57 %	10,51 %	8,33	10,47 %
4	1	64,11 %	11,29 %	10,23 %	8,08	10,44 %
5	1	62,30 %	10,98 %	9,89 %	7,77	10,40 %
6	1	61,56 %	11,04 %	9,89 %	7,83	10,34 %
7	1	60,87 %	11,24 %	9,95 %	8,08	10,16 %
0	2	60,85 %	9,45 %	8,86 %	6,58	10,38 %
1	2	60,85 %	9,45 %	8,86 %	6,58	10,38 %
2	2	60,85 %	9,48 %	8,88 %	6,6	10,38 %
3	2	60,77 %	9,51 %	8,89 %	6,64	10,37 %
4	2	60,32 %	9,45 %	8,84 %	6,58	10,36 %
5	2	59,50 %	9,42 %	8,76 %	6,54	10,34 %
6	2	59,05 %	9,47 %	8,75 %	6,6	10,28 %
7	2	58,49 %	9,71 %	8,78 %	6,87	10,11 %
0	3	56,93 %	8,22 %	7,81 %	5,58	10,26 %
1	3	56,93 %	8,22 %	7,81 %	5,58	10,26 %
2	3	56,93 %	8,22 %	7,81 %	5,58	10,26 %
3	3	56,93 %	8,25 %	7,83 %	5,61	10,25 %
4	3	56,89 %	8,29 %	7,87 %	5,66	10,25 %
5	3	56,75 %	8,35 %	7,82 %	5,72	10,24 %
6	3	56,58 %	8,39 %	7,84 %	5,76	10,19 %
7	3	56,20 %	8,60 %	7,91 %	5,98	10,03 %
0	4	53,66 %	6,76 %	6,55 %	4,24	10,17 %
1	4	53,66 %	6,76 %	6,55 %	4,24	10,17 %
2	4	53,66 %	6,76 %	6,55 %	4,24	10,17 %
3	4	53,66 %	6,77 %	6,55 %	4,24	10,17 %
4	4	53,63 %	6,78 %	6,56 %	4,25	10,16 %
5	4	53,63 %	6,78 %	6,56 %	4,26	10,15 %
6	4	53,63 %	6,84 %	6,61 %	4,32	10,11 %

Drugi pristop

k1	k2	$\overline{T^+/B}$	$\overline{T^+/A}$	$\overline{T^+/(A \cup B)}$	Dobiček	$\overline{A/\mathcal{B}}$
0	0	86,79 %	0,76 %	0,76 %	−3,91	94,07 %
1	0	86,79 %	0,76 %	0,76 %	−3,91	94,07 %
2	0	86,79 %	0,76 %	0,76 %	−3,91	94,07 %
3	0	86,79 %	0,76 %	0,76 %	−3,91	94,07 %
4	0	86,79 %	0,76 %	0,76 %	−3,91	94,07 %
5	0	86,79 %	0,76 %	0,76 %	−3,91	94,07 %
6	0	86,79 %	0,76 %	0,76 %	−3,91	94,07 %
7	0	86,30 %	0,76 %	0,76 %	−3,91	94,07 %
0	1	86,79 %	0,76 %	0,76 %	−3,91	94,07 %
1	1	86,04 %	22,39 %	22,25 %	19,17	10,55 %
2	1	86,04 %	22,39 %	22,25 %	19,17	10,55 %
3	1	86,04 %	22,39 %	22,25 %	19,17	10,55 %
4	1	86,04 %	22,39 %	22,25 %	19,17	10,55 %
5	1	86,04 %	22,39 %	22,25 %	19,17	10,55 %
6	1	86,04 %	22,39 %	22,25 %	19,17	10,55 %
7	1	85,55 %	22,31 %	22,15 %	19,09	10,55 %
0	2	76,05 %	0,75 %	0,75 %	−3,91	94,07 %
1	2	75,30 %	15,11 %	15,09 %	11,84	10,55 %
2	2	74,44 %	14,67 %	14,65 %	11,45	10,53 %
3	2	73,94 %	14,70 %	14,67 %	11,51	10,51 %
4	2	73,08 %	14,44 %	14,41 %	11,29	10,49 %
5	2	72,09 %	14,25 %	14,22 %	11,08	10,48 %
6	2	71,80 %	14,30 %	14,26 %	11,14	10,47 %
7	2	71,18 %	14,26 %	14,21 %	11,14	10,46 %
0	3	76,05 %	0,75 %	0,75 %	−3,91	94,07 %
1	3	75,30 %	15,11 %	15,09 %	11,84	10,55 %
2	3	74,44 %	14,69 %	14,67 %	11,47	10,53 %
3	3	73,86 %	14,74 %	14,69 %	11,55	10,51 %
4	3	72,59 %	14,44 %	14,41 %	11,29	10,48 %
5	3	70,92 %	14,18 %	14,14 %	11,01	10,45 %
6	3	70,35 %	14,22 %	14,15 %	11,05	10,43 %
7	3	69,54 %	14,27 %	14,11 %	11,15	10,41 %
0	4	76,05 %	0,75 %	0,75 %	−3,91	94,07 %
1	4	75,30 %	15,11 %	15,09 %	11,84	10,55 %
2	4	74,44 %	14,69 %	14,67 %	11,47	10,53 %
3	4	73,86 %	14,75 %	14,70 %	11,56	10,50 %
4	4	72,59 %	14,47 %	14,43 %	11,32	10,48 %
5	4	70,77 %	14,21 %	14,14 %	11,04	10,45 %
6	4	70,04 %	14,24 %	14,13 %	11,08	10,43 %

Dodatek B

Rezultati napovedovanja izdelkov

B.1 Primerjalne metode

Naključna metoda

k	$\overline{T^+/I}$	$\overline{T^+/C}$	$\overline{T^+/(C \cup I)}$	$\overline{C/\mathcal{I}}$
100	100,00 %	0,54 %	0,54 %	100,00 %
93	93,28 %	0,54 %	0,54 %	92,88 %
86	86,48 %	0,54 %	0,54 %	85,94 %
80	80,08 %	0,54 %	0,54 %	79,86 %
73	72,99 %	0,54 %	0,54 %	72,92 %
66	66,26 %	0,54 %	0,54 %	65,97 %
60	59,87 %	0,53 %	0,53 %	59,90 %
53	51,78 %	0,53 %	0,53 %	52,95 %
46	45,35 %	0,53 %	0,53 %	45,83 %
40	40,48 %	0,55 %	0,54 %	39,93 %
33	32,99 %	0,53 %	0,53 %	32,99 %
26	26,37 %	0,54 %	0,53 %	25,87 %
20	19,77 %	0,53 %	0,52 %	19,97 %
13	12,99 %	0,54 %	0,52 %	12,85 %
6	6,07 %	0,54 %	0,50 %	5,90 %
0	0,00 %	0,00 %	0,00 %	0,00 %

Metoda kategorij

k	$\overline{T^+/I}$	$\overline{T^+/C}$	$\overline{T^+/(C \cup I)}$	$\overline{C/\mathcal{I}}$
0	0,00 %	0,00 %	0,00 %	0,00 %
1	2,63 %	6,25 %	2,63 %	0,17 %
2	4,43 %	5,40 %	2,90 %	0,35 %
3	7,22 %	6,05 %	3,71 %	0,52 %
4	8,83 %	5,80 %	3,83 %	0,69 %
5	10,35 %	5,39 %	3,82 %	0,87 %
6	12,37 %	5,50 %	4,02 %	1,04 %
7	13,23 %	5,21 %	3,97 %	1,22 %
8	15,81 %	5,37 %	4,25 %	1,39 %
9	16,89 %	5,09 %	4,12 %	1,56 %
10	18,23 %	5,02 %	4,13 %	1,74 %
11	18,93 %	4,76 %	3,99 %	1,91 %
12	21,19 %	4,81 %	4,09 %	2,08 %
13	22,09 %	4,74 %	4,09 %	2,26 %
14	22,99 %	4,58 %	3,98 %	2,43 %
15	24,04 %	4,54 %	3,99 %	2,60 %

Linearna regresija

k	$\overline{T^+/I}$	$\overline{T^+/C}$	$\overline{T^+/(C \cup I)}$	$\overline{C/\mathcal{I}}$
0	92,23 %	1,18 %	1,18 %	53,02 %
0,1	83,20 %	25,70 %	25,00 %	3,87 %
0,1	73,62 %	31,46 %	29,81 %	2,95 %
0,2	64,03 %	35,55 %	32,27 %	2,33 %
0,3	55,54 %	38,60 %	33,07 %	1,81 %
0,3	49,31 %	40,50 %	32,91 %	1,48 %
0,4	44,55 %	41,01 %	31,97 %	1,32 %
0,5	41,00 %	40,61 %	31,02 %	1,21 %
0,5	38,12 %	39,58 %	30,01 %	1,11 %
0,6	35,69 %	39,09 %	29,07 %	1,05 %
0,7	33,67 %	38,11 %	28,16 %	1,00 %
0,7	31,87 %	37,10 %	27,20 %	0,95 %
0,8	30,34 %	35,99 %	26,23 %	0,92 %
0,9	28,88 %	34,82 %	25,31 %	0,89 %
0,9	27,52 %	33,55 %	24,34 %	0,87 %
1	26,23 %	32,25 %	23,41 %	0,86 %

KNN regresija

k	$\overline{T^+/I}$	$\overline{T^+/C}$	$\overline{T^+/(C \cup I)}$	$\overline{C/x}$
0	100,00 %	0,54 %	0,54 %	100,00 %
0,1	9,54 %	11,58 %	5,50 %	2,26 %
0,1	9,54 %	11,58 %	5,50 %	2,26 %
0,2	9,54 %	11,58 %	5,50 %	2,26 %
0,3	9,54 %	11,58 %	5,50 %	2,26 %
0,3	9,54 %	11,58 %	5,50 %	2,26 %
0,4	5,15 %	6,33 %	2,83 %	1,73 %
0,5	5,15 %	6,33 %	2,83 %	1,73 %
0,5	5,15 %	6,33 %	2,83 %	1,73 %
0,6	5,15 %	6,33 %	2,83 %	1,73 %
0,7	5,15 %	6,33 %	2,83 %	1,73 %
0,7	3,55 %	4,14 %	1,75 %	1,48 %
0,8	3,55 %	4,14 %	1,75 %	1,48 %
0,9	3,55 %	4,14 %	1,75 %	1,48 %
0,9	3,55 %	4,14 %	1,75 %	1,48 %
1	3,55 %	4,14 %	1,75 %	1,48 %

Regresija naključnih gozdov

k	$\overline{T^+/I}$	$\overline{T^+/C}$	$\overline{T^+/(C \cup I)}$	$\overline{C/x}$
0	100,00 %	0,54 %	0,54 %	100,00 %
0,1	63,91 %	23,87 %	21,43 %	3,87 %
0,1	53,85 %	27,67 %	23,72 %	3,18 %
0,2	53,85 %	27,67 %	23,72 %	3,18 %
0,3	44,39 %	30,09 %	24,01 %	2,74 %
0,3	36,88 %	31,67 %	22,95 %	2,47 %
0,4	36,85 %	31,66 %	22,95 %	2,46 %
0,5	30,01 %	31,29 %	20,52 %	2,21 %
0,5	25,06 %	29,53 %	17,93 %	2,05 %
0,6	25,05 %	29,52 %	17,93 %	2,04 %
0,7	20,56 %	26,65 %	15,23 %	1,86 %
0,7	17,15 %	23,05 %	12,77 %	1,74 %
0,8	17,15 %	23,05 %	12,77 %	1,74 %
0,9	14,13 %	19,24 %	10,46 %	1,60 %
0,9	12,06 %	16,40 %	8,85 %	1,51 %
1	12,06 %	16,40 %	8,85 %	1,50 %

B.2 Pristop 1: Skupni sosedje

Utež 1

k	$\overline{T^+/I}$	$\overline{T^+/C}$	$\overline{T^+/(C \cup I)}$	$\overline{C/I}$
0	47,16 %	0,33 %	0,33 %	94,41 %
1	45,85 %	8,08 %	7,65 %	5,82 %
2	44,88 %	8,56 %	7,86 %	5,39 %
3	43,97 %	8,79 %	7,98 %	5,02 %
4	42,69 %	8,78 %	7,95 %	4,75 %
5	41,88 %	8,81 %	7,95 %	4,52 %
6	41,31 %	8,97 %	8,13 %	4,34 %
7	40,60 %	9,21 %	8,28 %	4,11 %
8	39,85 %	9,42 %	8,38 %	3,93 %
9	38,64 %	9,48 %	8,36 %	3,75 %
10	37,76 %	9,40 %	8,27 %	3,63 %
11	37,17 %	9,52 %	8,28 %	3,51 %
12	36,50 %	9,59 %	8,30 %	3,38 %
20	31,33 %	9,26 %	7,73 %	2,69 %
28	27,68 %	9,74 %	7,60 %	2,19 %
36	24,47 %	9,17 %	7,42 %	1,75 %
44	21,55 %	9,43 %	7,37 %	1,43 %
52	18,95 %	10,34 %	7,43 %	1,17 %
60	15,79 %	9,65 %	6,44 %	0,99 %
68	12,53 %	7,31 %	4,98 %	0,83 %
76	10,28 %	6,87 %	4,48 %	0,70 %
84	9,82 %	6,42 %	4,39 %	0,60 %
92	9,90 %	6,88 %	4,59 %	0,53 %
100	10,37 %	7,39 %	4,81 %	0,51 %

Utež 2

k	$\overline{T^+/I}$	$\overline{T^+/C}$	$\overline{T^+/(C \cup I)}$	$\overline{C/I}$
0	47,16 %	0,33 %	0,33 %	94,41 %
1	45,85 %	8,08 %	7,65 %	5,82 %
2	45,06 %	8,45 %	7,84 %	5,46 %
3	44,34 %	8,68 %	7,96 %	5,18 %
4	43,23 %	8,76 %	7,99 %	4,96 %
5	42,52 %	8,83 %	7,96 %	4,79 %
6	42,11 %	8,91 %	8,04 %	4,64 %
7	41,52 %	9,18 %	8,26 %	4,45 %
8	40,79 %	9,34 %	8,33 %	4,33 %
9	40,34 %	9,42 %	8,44 %	4,21 %
10	39,80 %	9,48 %	8,45 %	4,11 %

11	39,30 %	9,52 %	8,45 %	4,03 %
12	38,53 %	9,57 %	8,44 %	3,91 %
20	33,86 %	10,03 %	8,33 %	3,27 %
28	30,16 %	10,19 %	7,95 %	2,81 %
36	26,64 %	9,24 %	7,38 %	2,41 %
44	25,28 %	9,93 %	8,04 %	2,09 %
52	22,45 %	10,13 %	7,85 %	1,81 %
60	19,19 %	10,52 %	7,24 %	1,58 %
68	16,59 %	9,71 %	6,51 %	1,36 %
76	14,31 %	9,16 %	5,89 %	1,22 %
84	14,12 %	9,05 %	6,04 %	1,12 %
92	13,98 %	9,91 %	6,27 %	1,04 %
100	12,78 %	9,89 %	5,90 %	0,96 %

B.3 Pristop 2: Vsi sosedje

Utež 1

k	$\overline{T^+/I}$	$\overline{T^+/C}$	$\overline{T^+/(C \cup I)}$	$\overline{C/\mathcal{I}}$
0	87,13 %	0,53 %	0,53 %	95,81 %
12	80,87 %	8,28 %	8,10 %	6,92 %
25	74,93 %	8,69 %	8,45 %	6,02 %
37	70,50 %	9,35 %	8,94 %	5,34 %
50	66,35 %	9,42 %	9,01 %	4,91 %
62	62,23 %	9,78 %	8,68 %	4,56 %
75	59,35 %	8,84 %	8,35 %	4,16 %
87	58,10 %	9,21 %	8,55 %	3,87 %
100	56,55 %	9,87 %	8,85 %	3,61 %
112	54,18 %	10,42 %	8,82 %	3,36 %
125	50,92 %	9,31 %	8,07 %	3,04 %
137	48,75 %	9,49 %	8,23 %	2,78 %
150	43,06 %	6,88 %	6,72 %	2,42 %
162	39,14 %	6,96 %	6,71 %	2,14 %
175	36,64 %	7,36 %	7,02 %	1,86 %
187	33,20 %	7,53 %	7,06 %	1,63 %
200	26,90 %	7,81 %	6,98 %	1,27 %
212	21,91 %	7,88 %	6,72 %	1,04 %
225	17,02 %	8,19 %	6,37 %	0,76 %
237	13,78 %	8,83 %	6,19 %	0,56 %
250	6,22 %	8,74 %	4,13 %	0,24 %
262	1,88 %	4,25 %	1,57 %	0,10 %
275	0,00 %	0,00 %	0,00 %	0,00 %
287	0,00 %	0,00 %	0,00 %	0,00 %
300	0,00 %	0,00 %	0,00 %	0,00 %

Utež 2

k	$\overline{T^+/I}$	$\overline{T^+/C}$	$\overline{T^+/(C \cup I)}$	$\overline{C/x}$
0	87,13 %	0,53 %	0,53 %	95,81 %
33	82,18 %	8,35 %	8,26 %	6,85 %
66	78,31 %	9,25 %	8,95 %	5,94 %
100	73,01 %	10,20 %	9,66 %	5,03 %
133	68,59 %	10,40 %	9,93 %	4,39 %
166	64,31 %	11,03 %	10,03 %	3,92 %
200	60,03 %	11,54 %	10,38 %	3,38 %
233	55,14 %	12,27 %	10,88 %	2,82 %
266	45,52 %	12,73 %	10,86 %	2,11 %
300	35,76 %	12,00 %	9,07 %	1,70 %
333	31,25 %	8,72 %	7,46 %	1,48 %
366	28,69 %	8,45 %	7,19 %	1,31 %
400	27,06 %	9,28 %	7,72 %	1,16 %
433	24,44 %	10,05 %	7,87 %	0,97 %
466	23,38 %	10,16 %	7,90 %	0,89 %
500	21,65 %	10,25 %	7,62 %	0,84 %
533	20,42 %	10,23 %	7,35 %	0,77 %
566	19,69 %	10,33 %	7,32 %	0,73 %
600	17,42 %	10,02 %	6,67 %	0,65 %
633	16,63 %	10,27 %	6,71 %	0,60 %
666	14,49 %	7,55 %	5,10 %	0,57 %
700	13,83 %	7,67 %	5,00 %	0,54 %
733	13,80 %	7,79 %	5,10 %	0,52 %
766	13,07 %	7,85 %	5,13 %	0,46 %
800	12,61 %	8,03 %	5,18 %	0,43 %

Literatura

[1] S. O'Connor, "Amazon unpacked", *Financial Times Magazine*, 2013. Dostopno na: https://www.ft.com/content/ed6a985c-70bd-11e2-85d0-00144feab49a#slide0 [20.09.2017]

[2] J. Rowley, "Promotion and marketing communications in the information marketplace", *Library review*, vol. 47, no. 8, pp. 383–387, 1998.

[3] Capital One Financial Corporation, *10 of the Best Sales Promotions of all Time*, 2014, (e-book). Dostopno na: https://www.sparkpay.com/top-10-sales-promotions [28.08.2017]

[4] E. Weisberg. (2015) 9 sales promotion examples. Dostopno na: https://thrivehive.com/sales-promotion-examples [20.09.2017]

[5] R. C. Blattberg and S. A. Neslin, "Sales promotion models", *Handbooks in Operations Research and Management Science*, vol. 5, pp. 553–609, 1993.

[6] Walker Sands Communications. (2014) Walker sands' 2014 future of retail study. Dostopno na: https://www.walkersands.com/2014-futureofretail [26.08.2017]

[7] A. Maybank and A. Wilson, *By Invitation Only: How We Built Gilt and Changed the Way Millions Shop*. Penguin Publishing Group, 2012.

[8] S. Ingo and K. Ivan, "Adaptation to drifting user's interests", *In proceedings of ECML2000 Workshop: Machine Learning in New Information Age*, 2000.

[9] K. S. Moorthy and I. P. Png, "Market segmentation, cannibalization, and the timing of product introductions", *Management Science*, vol. 38, no. 3, pp. 345–359, 1992.

[10] J. A. McCarty and M. Hastak, "Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression", *Journal of Business Research*, vol. 60, no. 6, pp. 656–662, 2007.

[11] K. K. Tsiptsis and A. Chorianopoulos, *Data mining techniques in CRM: inside customer segmentation.* John Wiley & Sons, 2011.

[12] M. Volpi. (2014) 5 tips for avoiding promotion fatigue & regret. Dostopno na: https://gocatalant.com/blog/5-tips-for-avoiding-promotion-fatigue-regret [20.09.2017]

[13] A. Y. L. Chong, B. Li, E. W. Ngai, E. Ch'ng, and F. Lee, "Predicting online product sales via online reviews, sentiments, and promotion strategies: A big data architecture and neural network approach", *International Journal of Operations & Production Management*, vol. 36, no. 4, pp. 358–383, 2016.

[14] F. Halper, "Predictive analytics for business advantage", *TDWI Research*, 2014.

[15] J. F. Hair, "Knowledge creation in marketing: the role of predictive analytics", *European Business Review*, vol. 19, no. 4, pp. 303–315, 2007.

[16] J. Bowden. (2014) How to use predictive analytics. Dostopno na: http://www.digital-warriors.com/use-predictive-analytics/?utm_source=rss&utm_medium=rss&utm_campaign=use-predictive-analytics [18.09.2017]

[17] Futurisimo d.o.o. (2011) Kuponko.si. Dostopno na: https://www.kuponko.si [26.08.2017]

[18] Uber Zon Club. (2015) Uberzonclub.com. Dostopno na: http://www.uberzonclub.com [26.08.2017]

[19] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms, Third Edition*, 3rd ed. The MIT Press, 2009.

[20] M. E. Newman, "Analysis of weighted networks", *Physical review E*, vol. 70, no. 5, p. 056131, 2004.

[21] The Dobney Corporation Limited. Database analysis to big data analysis. Dostopno na: http://www.dobney.com/Intelligence/database.htm [18.09.2017]

[22] S. Zhang, C. Zhang, and Q. Yang, "Data preparation for data mining", *Applied Artificial Intelligence*, vol. 17, no. 5-6, pp. 375–381, 2003.

[23] Wikipedia. (2017) Data analysis. Dostopno na: https://en.wikipedia.org/w/index.php?title=Data_analysis&oldid=801957018 [23.09.2017]

[24] S. Morgenthaler, "Exploratory data analysis", *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1, no. 1, pp. 33–44, 2009.

[25] O. Corporation. (2005) MySQL Workbench. Dostopno na: https://www.mysql.com/products/workbench [16.09.2017]

[26] Amazon.com, Inc. (2017) Product advertising api. Dostopno na: http://docs.aws.amazon.com/AWSECommerceService/latest/DG [25.08.2017]

[27] Wikipedia. (2017) Regression analysis. Dostopno na: https://en.wikipedia.org/w/index.php?title=Regression_analysis&oldid=799445704 [08.09.2017]

[28] D. Montgomery, *Introduction to linear regression analysis.* Hoboken, NJ: Wiley, 2012.

[29] L. Breiman, "Random forests", *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[30] Python Software Foundation. (2008) Python. Dostopno na: https://www.python.org/ [01.09.2017]

[31] N. Janko. (2017) Algorithmic optimization of sales acceleration. Dostopno na: https://github.com/nikolai5slo/algo_sales_acceleration [25.09.2017]

[32] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, 2016.